What's the Point: Semantic Segmentation with Point Supervision

Amy Bearman[1] Olga Russakovsky[2] Vittorio Ferrari[3] Li Fei-Fei[1]

Contributions

Goal: Obtain the most annotation cost-effective supervision for semantic image segmentation.

- Novel, cost-efficient **supervision regime** for semantic segmentation based on humans pointing to objects.
- Extensive human study to collect **point annotations** for PASCAL VOC 2012, and released **annotation interfaces**.
- A generic **objectness prior** incorporated directly in the loss to guide the training of a CNN.

Novel supervision regime

Problem: Assign one class label to every pixel in an image.

- **Training**: Standard regime = costly per-pixel annotations

- **Levels of supervision**

full supervision | image-level labels | points | squiggles

- **Key insight**: Annotating **one pixel** per training image significantly improves **segmentation annotation** and only marginally increases the **annotation cost** as compared to image-level labels.
- **Loss function for point-level supervision**: We have a small set of supervised pixels, and other pixels just belong to some class in L.

$$\mathcal{L}_{point}(S, G, L, L') = -\frac{1}{|L|}\sum_{c\in L}\log\left(S_{t_c c}\right) - \frac{1}{|L'|}\sum_{c\in L'}\log\left(1 - S_{t_c c}\right) - \sum_{i\in I_s}\alpha_i \log\left(S_{iG_i}\right)$$

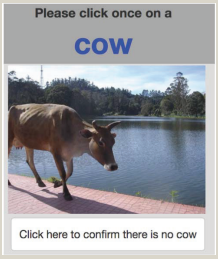
Ground truth map; Classes not in image; Encourages each class in L to have high probability on ≥ 1 pixel in the image [Pathak 2015]; No pixels should have high probability for classes not present in the image; Multinomial logistic loss on softmax probabilities, for supervised pixels

3D per-pixel map of softmax probabilities for the image of size W x H with N classes; Classes in image; Softmax probability of class c at pixel t_c, the highest scoring pixel for that class; Relative importance of each supervised pixel; Ground truth class of i-th pixel

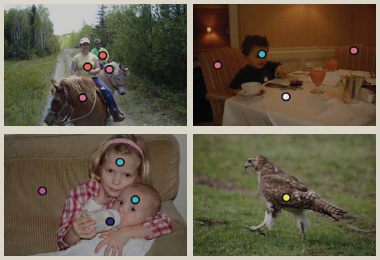
- **Model**: Fully convolutional network [Long 2015].

Crowdsourcing point annotations

AMT annotation UI



Example points collected



Measuring the annotation times:

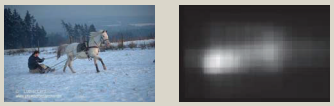
- Points and squiggles: measured directly during data collection.
- Other types of supervision: we rely on times from literature.

Reported annotation times:

- Image-level labels: **20.0 sec/image**
- Points: **22.1 sec/image**
- Squiggles: **34.9 sec/image**
- Full supervision: **239.7 sec/image**

Objectness prior in CNN loss

Purpose of the objectness prior: Helps correctly infer the spatial extent of objects for models trained with very few supervised pixels.



original image | objectness prior

- **Obtaining the prior**: Assign each pixel the average objectness score of all windows containing it. Scores are obtained from the model of [Alexe 2012], which is trained on 50 images from datasets that do not overlap with PASCAL VOC 2012.
- **Incorporation into loss function**: Provides a probability for whether a pixel is in the set of all object classes (O), instead of background.

$$\mathcal{L}_{obj}(S, P) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left(P_i \log\left(\sum_{c\in\mathcal{O}} S_{ic}\right) + (1 - P_i)\log\left(1 - \sum_{c\in\mathcal{O}} S_{ic}\right)\right)$$

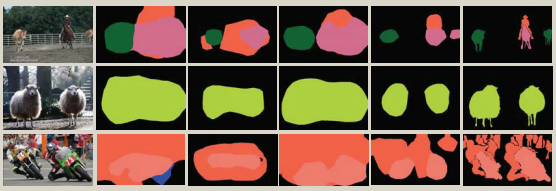
Per-pixel map of objectness probabilities for the image; Probability that pixel i belongs to an object; Probability that pixel i belongs to background

3D per-pixel map of softmax probabilities for the image of size W x H; Set of pixels in image; Set of object classes; Softmax probability of class c at pixel i

Results on PASCAL VOC 2012 dataset [Everingham 2010]

- **Effects of point supervision + objectness**: The combined effect results in a **+13% mIOU** over image-level labels.

Supervision	Time (s)	mIOU (%)
Image-level	20.0	29.8
Image-level + objectness	20.3	32.2
1Point	22.1	35.1
1Point + objectness	22.4	42.7



original image | image-level supervision | image-level + objectness | point-level supervision | point-level + objectness | ground truth

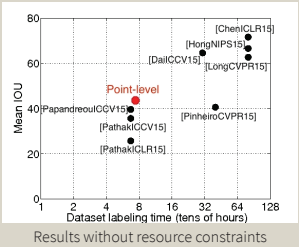
- **Point supervision variations**: Multiple object instances and multiple annotators achieve only modest improvements over single points.

Supervision	Time (s)	mIOU (%)
1Point	22.4	42.7
1Point (random annotators)	22.4	42.8 - 43.8
1Point (3 annotators)	29.6	43.8
All Instances	23.6	42.7
All Instances (weighted)	23.5	43.4
1Point (random points)	240	46.1

- **Segmentation on an annotation budget**: Point supervision provides the best trade-off between annotation time and segmentation accuracy.

Supervision	mIOU (%)
Full (883 imgs)	22.1
Image-level (10,582 imgs)	29.8
Squiggle-level (6,064 imgs)	40.2
Point-level (9,576 imgs)	42.9

Accuracy of models on the PASCAL VOC 2012 validation set given a fixed annotation budget.



Results without resource constraints on the PASCAL VOC 2012 test set.

Bibliography

- J. Long, et al. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.
- B. Alexe, et al. Measuring the objectness of image windows. PAMI 2012.
- D. Pathak, et al. Constrained Convolutional Neural Networks for Weakly Supervised Segmentation. ICCV 2015.
- M. Everingham, et al. The Pascal Visual Object Classes (VOC) challenge. 2010.

[1]Stanford University [2]Carnegie Mellon University [3]University of Edinburgh