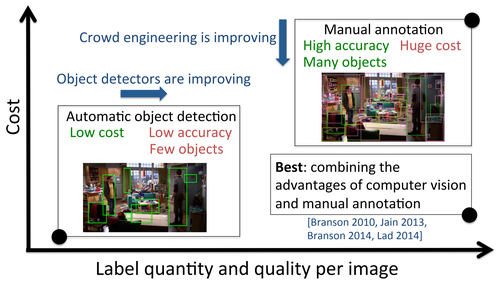
Best of both worlds: human-machine collaboration for object annotation

Olga Russakovsky[1] Li-Jia Li[2] Li Fei-Fei[1]

[1] Stanford University [2] Snapchat (this work was done at Yahoo! Labs)

Introduction

Goal: To build a principled framework for accurately and efficiently localizing objects in an image



Summary: We developed a principled human-in-the-loop framework integrating state-of-the-art scene understanding models with state-of-the-art crowd engineering methods for detecting objects in images.

Key differences with prior work:

1. Complex, open-ended annotation task
2. Novel Markov Decision Process formulation
3. Using *multiple* types of human input and *multiple* types of computer vision models
4. Automatically trading off utility, precision, and cost of annotation

Computer vision conclusions:

- Current detectors recognize 1-2 objects per image
- Our system can collect large-scale datasets

Crowd engineering conclusions:

- Good to principally combine multiple human tasks
- Asking complex questions may be more efficient

Method

Input: image & constraints

A) Image to label:

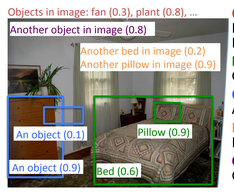
B) Constraints:

1. utility:
   Given a function f (Region,Class) -> [0, 1] indicating the importance/usefulness of a label (R,C), the utility of a labeling $(R_1,C_1)\ldots(R_N,C_N)$ is $\Sigma f(R_i,C_i)$
2. precision:
   If returned N detections and M are correct then precision is M/N
3. and/or budget
   Budget is human annotation time

Compute image beliefs

Image annotation state

computer vision (+ human input)



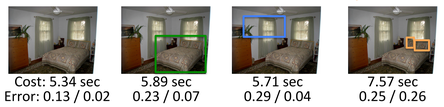
Classifiers: Krizhevsky 2012, Hoffman 2014, Caffe
Detectors: Girshick 2014
Objectness: Alexe 2012
Instances statistics: ILSVRC2014
Objects statistics: Greene 2014

Select optimal question within budget

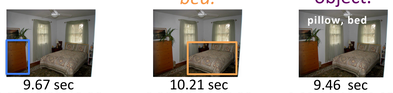
Update image beliefs

Return optimal annotation

Human tasks: 4 binary tasks, 3 open-ended tasks



1: Is there a *fan*?	2: Is this a *bed*?	3: Is this an object?	4: Are there more *pillows*?
Cost: 5.34 sec	5.89 sec	5.71 sec	7.57 sec
Error: 0.13 / 0.02	0.23 / 0.07	0.29 / 0.04	0.25 / 0.26



5: Name this object.	6: Draw another *bed*.	7: Name another object.
9.67 sec	10.21 sec	9.46 sec
0.25 / 0.08 / 0.06	0.28 / 0.16 / 0.29	0.02 / 0.12 / 0.05

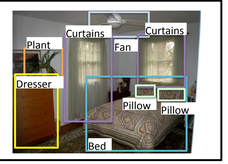
Cost and error rates
Empirical from Mturk
Errors rates are:
False Negative /
False Positive /
(Wrong on Attempt)

Note for future work: Distinguishing between "good" and "bad" bounding box is very difficult for humans

Output: final annotation

within specified precision, utility, and/or budget

Given N detections (Region, Class, Probability) $(R_1,C_1,p_1)\ldots(R_N,C_N,p_N)$:
Expected utility is $\Sigma p_i f(R_i,C_i)$
Expected precision is $\Sigma p_i/N$



Select question: Markov Decision Process

State: set of image descriptions, with probabilities
Action: a question to ask humans
Reward: increase in estimated utility of labeling divided by cost
Transition probabilities correspond to expected user answer. Computed from:

1. Current estimate of correct response
2. Pre-computed human error rates

Optimization: 2-step lookahead search

Update image beliefs: Joint computer vision and user model

[Branson 2010]

$$\underbrace{P(E|I,\mathcal{U})}_{\text{Annotation probability}} \propto \underbrace{P(E|I)}_{\text{Computer vision model}} \prod_k \underbrace{P(u_k|E)}_{\text{User model}}$$

Event (e.g., "there is a cat in this box") — Image — User feedback — Number of users — User accuracy or error rate

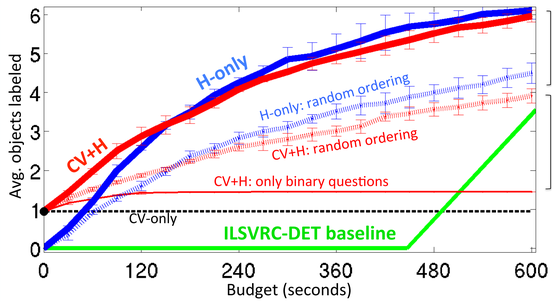
Our challenge: computing $P(u_k \mid E)$ when E = "there is a cat in this box" and u_k is a response to a previous (potentially unrelated) question e.g., "is there a cow in this image?"

Results

Data: ILSVRC 2014-DET validation set, 200 object classes, 2216 images (≥4 instances per image)
Models: [Krizhevsky 2012, Hoffman 2014, Caffe, Girshick 2014. Alexe 2012, Greene 2014

Quantitative results

(1) Computer vision (CV) and human input (H) are mutually beneficial (at low budget)

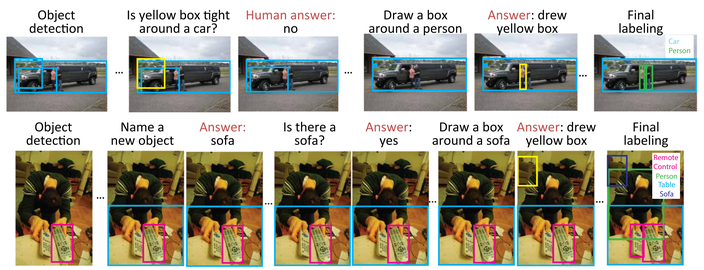


(2) MDP is effective at selecting tasks

(3) Complex human tasks are necessary

(4) Our strategy is more effective than ILSVRC baseline

Qualitative results



Bibliography

Partially sponsored by ONR MURI, by Yahoo! Labs, and by NVIDIA

[Alexe 2012] B. Alexe et al. Measuring the objectness of image windows. PAMI 2012.
[Branson 2010] S. Branson et al. Visual Recognition with Humans in the Loop. ECCV 2010.
[Branson 2014] S. Branson et al. Active annotation translation. CVPR 2014.
[Girshick 2014] R. Girshick et al. Rick feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014.
[Hoffman 2014] J. Hoffman et al. LSDA: Large scale detection through adaptation. NIPS 2014.
[Jain 2013] S. Jain and K. Grauman. Predicting sufficient annotation strength for interactive foreground segmentation. ICCV 2013.
[ILSVRC] O. Russakovsky. J. Deng et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.
[Lad 2014] S. Lad and D. Parikh. Interactively guiding semi-supervised clustering via attribute-based explanations. ECCV 2014.