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

Olga Russakovsky1 Li-Jia Li2 Li Fei-Fei1
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.

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

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.

Method

Input: image & constraints
A) Image to label:
1. utility:
   - Given a function $f \in \{\text{Region}, \text{Class} \rightarrow 0 \text{ or } 1\}$ indicating the importance/score of a label $(R,C)$, the utility of a labeling $(R,C) - |(R,C)| \cup C(R,C)$
2. precision:
   - If returned $N$ detections and $M$ are correct then precision is $M/N$
3. and/or budget
   - Budget is human annotation time.

Image annotation state computer vision (+ human input)

Output: final annotation

Update human beliefs:

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

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

Select optimal question within budget

Human feedback

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

$p(E|I, U) \propto p(E|I) \prod_{k} p(U_k|E_k)$

|Event | $p(U_k|E_k)$ | $p(E|I)$ | User feedback | Number of users | User accuracy or error rate

| "Is there a cat in this box?" | 0.25 | 0.01 | 0.12 | 2 | 0.13 / 0.22
| "Is there a cow in this image?" | 0.29 | 0.24 | 0.15 | 2 | 0.15 / 0.26

Bibliography

Partially sponsored by NSF/CNS, Yahoo! Labs, and CDC 2010.

Results

Data: ILSVRC 2014-DET validation set, 200 object classes, 2216 images (4 images per image).

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

Verify-box: Is the yellow box around a person?

Draw-box: Draw a box around a person.

Final labeling

Table. Number of images, counts of human inputs, and error rates for crowdsourced labeling.

User answer.

Answer: Yes

Answer: No

Answer: Yellow box below