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 while at Yahoo! Labs).

The long-standing goal of localizing every object in an image remains elusive. Manually annotating objects is quite expensive despite crowd engineering innovations. Current automatic object detectors can accurately detect at most a few objects per image. This paper brings together the latest advancements in object detection and in crowd engineering into a principled framework for accurately and efficiently localizing objects in images.

The input to the system is an image to annotate and a set of annotation constraints: (1) desired **utility** of labeling, which is a generalization of the number of labeled objects, (2) desired **precision** of the labeling and/or (3) the **budget**, which is the human cost of the labeling. Our system automatically solicits feedback from human workers ("users") to annotate the image subject to these constraints, as illustrated in Figure 1. The output is a set of object annotations, informed by humans and computer vision.

One important decision is which questions to pose to the human labelers. In computer vision with human-in-the-loop approaches, human intervention has ranged from binary question-and-answer [1] to attribute-based feedback [4] to free-form object annotation [6]. Binary questions are not sufficient in our setting. Asking users to draw bounding boxes is expensive: obtaining an accurate box takes between 7 seconds [2] to 42 seconds [5], and with 23 objects in an average indoor scene the cost quickly adds up. Based on insights from [3, 5], it is best to use a variety of human interventions.

Model. The main component of our approach is automatically selecting the best question to ask. We quantify the tradeoff between cost and accuracy of annotation by formulating it as a Markov decision process (**MDP**). An MDP consists of states $\mathcal{S}$, actions $\mathcal{A}$, transition probabilities $\mathcal{P}$, and rewards $\mathcal{R}$.

States. At each time period, the MDP is in some state $S \in \mathcal{S}$. In our case, a state S is our set of current beliefs about the image I, computed by combining computer vision models with user input.

Actions. The MDP takes an action $a \in \mathcal{A}$ from state s, which transitions to state s' with probability $\mathcal{P}(s'|s,a)$. In our setting, the set of actions $\mathcal{A}$ correspond to the set of human questions that the system can ask. We use 7 different human tasks of varying levels of complexity, illustrated in Figure 1.

Transition probabilities. As a result of an action a from state s, the system moves into a new state s'; in other words, the current beliefs about the image get updated by the addition of a new user response. Transition probabilities correspond to our expectations on the user response to the question a. Computing these probabilities is the most mathematically challenging part, and requires the use of multiple computer vision models: image classifiers, object detectors, statistics about object classes and instances in images.

Rewards. After going from state s to s' through action a, the agent receives a reward $\mathcal{R}_a(s,s')$. In our case, the reward is the predicted increase in utility between states s and s' divided by the cost of action a.

Results. We evaluate both the accuracy and cost of our proposed system on the task of labeling the ILSVRC object detection dataset [5]. After 30 seconds of human labeling, our MDP combining computer vision with human input is able to label 2.77x more objects in an image on average than the same model without computer vision. We demonstrate that (1) computer vision and human input are mutually beneficial, (2) an MDP is an effective model for selecting human tasks, (3) complex human tasks are necessary for effective annotation, and (4) our annotation strategy is more effective than the original ILSVRC detection annotation system [5]. We also discuss in detail the challenges of designing the variety of human tasks.

We conclude with several take-home messages. First, from the **computer vision perspective**, current object detectors are far from perfect and can only detect a couple of objects in an average image. Further, accuracy drops rapidly when a tighter bounding box (with IOU higher than 0.5) is required. Our work can be used for collecting large-scale datasets with minimal supervision to improve the current state-of-the-art object detectors; in turn, the improved models will make our system more effective.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

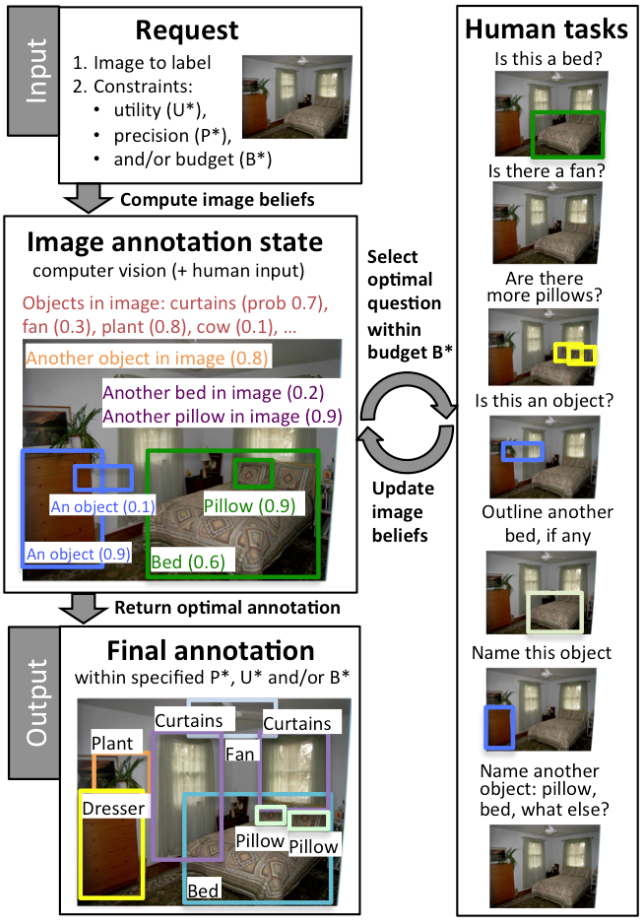

Figure 1: Overview of our system. Given a request for labeling an image, the system alternates between updating the image annotation and soliciting user feedback through human tasks. Upon satisfying the requester specifications, it terminates and returns a image with a set of bounding box annotations.

From a **crowd engineering perspective**, we demonstrated that it is worthwhile to combine multiple tasks in a principled framework. Our findings confirmed that asking slightly more complex tasks (such as immediately asking to draw the bounding box around an unannotated object instance rather than merely asking if one exists) is beneficial. This is in line with the findings of e.g., the COCO dataset curators [3] that asking slightly more complex human tasks (such as putting a dot on the object rather than merely asking if the object appears in the image) may be more efficient.

Finally, from an **application developer perspective**, we show that even though computer vision is not yet ready to detect all objects, we have a principled way of labeling all objects in a scene, trading off precision, utility and budget.

[1] Branson et al. Visual recognition with humans in the loop. *ECCV*, 2010.
[2] Jain and Grauman. Predicting sufficient annotation strength for interactive foreground segmentation. *ICCV*, 2013.
[3] Lin et al. Microsoft COCO: Common Objects in Context. 2014.
[4] Parkash and Parikh. Attributes for classifier feedback. *ECCV*, 2012.
[5] Russakovsky, Deng, et al. ImageNet Large Scale Visual Recognition Challenge. *CoRR*, abs/1409.0575, 2014.
[6] Vondrick, Patterson, and Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 2013.