Active Boundary Annotation using Random MAP Perturbations

Subhransu Maji
TTI Chicago

Tamir Hazan
University of Haifa, Israel

Tommi Jaakkola
MIT

Abstract

We address the problem of efficiently annotating labels of objects when they are structured. Often the distribution over labels can be described using a joint potential function over the labels for which sampling is provably hard but efficient maximum a-posteriori (MAP) solvers exist. In this setting we develop novel entropy bounds that are based on the expected amount of perturbation to the potential function that is needed to change MAP decisions. By reasoning about the entropy reduction and cost tradeoff, our algorithm actively selects the next annotation task. As an example of our framework we propose a boundary refinement task which can used to obtain pixel-accurate image boundaries much faster than traditional tools by focussing on parts of the image for refinement in a multi-scale manner.

1 Introduction

High quality data sets are important to develop novel approaches that can accurately solve challenging tasks. As current challenges become more and more complex, successful selection of approaches largely depends on the quality of annotations – a poor benchmark badly affects the evaluations and conclusions. Unfortunately, attaining high quality annotations for complex models, whether they describe objects in images, parses in sentences, or molecular structures in proteins, is a costly task as it involves an expert annotator to process each instance.

Active learning algorithms may interactively choose which data points to label in order to learn a labeling rule for all data points while substantially reducing the number of labels required [35]. Thus, it can be used to reduce the amount of time that is required by the expert to obtain high-quality annotations. Annotations for complex models are described by structured-labels, e.g., a sequence of labels that are strongly correlated. Specifically, image annotations provide a semantic label for each pixel. Unfortunately, conventional active learning approaches such as [10] assume the (pixels) labels are generated independently, thus they cannot be applied to achieve high-quality (image) annotations. These problems can be solved using Bayesian active learning, also known as Bayesian experimental design, since it relies on a probability model over all possible labels. Such an approach requires assessing the uncertainty of its probability model, a #P-hard problem in general [41], thus it is currently limited to simple probability models such as Gaussian processes [24, 18]. Alternatively, non-Bayesian approaches utilize the most likely or maximum a-posteriori (MAP) label rather than assessing the uncertainty of its probability model, a task that is substantially easier [44, 43, 36]. These approaches may leverage the recent effort that has gone into developing algorithms for recovering MAP assignments, either based on specific parametrized restrictions such as super-modularity [26] or by devising approximate methods based on linear programming relaxations [37]. However, MAP-based approaches are limited when describing uncertainties in users behavior and their annotations costs.

We propose a new uncertainty measure that can be efficiently computed by MAP perturbations. This measure provides a way to apply Bayesian active learning to high dimensional complex models while enjoying efficient MAP solvers such as graph-cuts or MPLP [3, 37]. While MAP perturbations have been considered recently [31, 25, 38, 14, 16, 15, 32, 30, 11], their relation to uncertainty measures has not. Specifically, we construct an upper bound to the entropy function using the expected amount of perturbation that is required to change the MAP decision. We show that this upper bound is an uncertainty measure, i.e., it is nonnegative, reaches its maximal value on the uniform distribution and its minimal value on the zero-one distribution. Our approach excels in cases where observations carry strong signals (local evidence), but are also guided by strong consistency constraints (couplings). This "high-signal, high-coupling" domain is typical in machine learning applications and easy for MAP-solvers. Nevertheless, it creates ragged energy landscapes and classical sampling methods for estimating uncertainties, such as

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

Markov chain Monte Carlo (MCMC) samplers, are provably hard [12, 46].

We begin by reviewing the previous work on interactive annotations and active learning approaches. We subsequently describe the Bayesian active learning approach, followed by our new upper bound for the entropy function that rely on measuring the boundaries of MAP decisions. Additionally, we describe how to deal with data dependent cost such as high curvature boundary annotation that is more expensive to annotate than straight lines. We conclude with real-life experiments on annotating high-resolution image boundaries showing significantly reductions in annotation cost over traditional approaches.

2 Previous work

Large data sets in computer vision often rely on crowdsourcing for obtaining annotations. Some examples include LabelMe [34] and ImageNet [9]. However quality control and cost effectiveness are important concerns in this setting. Our work differs as it suggests to use active learning to interact with the annotator in order to reduce the crowdsourcing costs of acquiring high quality annotations.

Theoretical aspects of active learning has been investigated in machine learning [10, 40, 8, 1, 13]. Generally, these works assume binary classification of the data, whose labels are independent and identically distributed (i.i.d.) according to some distribution. This setting does not fit to analyze pixel-wise image annotations, since the pixels labels of the same image introduce dependencies to the statistical process. Recently, active learning approaches were applied to complex structured labels, thus effectively using non-i.i.d. setting [33, 44, 4, 43, 36]. These approaches rely on efficient max-solvers that find the best structured label, or equivalently the MAP assignment. Unfortunately, these approaches do not allow to consider elaborate user behavior that is important to attain high quality annotations while minimizing their costs. In our work we use random perturbations to estimate the MAP decision boundaries in order to attain a better understanding of the uncertainties in the model.

Bayesian approaches to active learning are extensively applied to deal with uncertainties of models. Stemming from experimental design [5], Bayesian approaches consider a probability model over the all possible labels (e.g., annotations), and active learning is applied to design an experiment (e.g., annotation of sub-area of the image) that minimizes the expected loss of the probability model. [21, 39] use Dirichlet distributions and the entropy as their loss function. Alternatively, Gaussian processes have emerged as effective probability models for complex tasks. [6] use them and their variance as their loss function, while [24] use both their variance and their mean to measure their annotation cost, and [27, 18, 20] use their entropies to measure their loss. Obtaining high quality annotations requires more elaborate loss functions. In our work we suggest to use both the uncertainty of the probability model as well as the workload on the annotator. Recently, a few works consider the more elaborate setting of Bayesian active learning and crowdsourcing. [42] introduce loss functions that include model uncertainty as well as the annotation effort. Alternatively, [45] introduce loss functions that include model uncertainty and the annotator's ability. In contrast, we focus on annotations of structured-labels while measuring the uncertainty of probability distributions over all possible labels in an exponentially large space of possible annotations. In our work we devise a new uncertainty measure that can be efficiently computed using MAP solvers. Thus we are able to utilize the efficient graph-cuts and MPLP algorithms in Bayesian active learning.

Perturb-max probability models that are defined by MAP perturbations have been recently introduced by [31], their sampling and learning approaches were described by [38, 11] and their generalization properties in [25, 15]. Their relations to the Gibbs distributions and their normalization constant (the partition functions) were discussed in [16, 14, 30]. However, none of these works considered uncertainty measures using MAP perturbations. In this work we present new upper bounds on the entropy function that can be efficiently computed using MAP perturbations. We also show that these upper bounds are uncertainty measures, since they are: (i) nonnegative, (ii) attain their maximal value for the uniform distribution, and (iii) attain their minimal value for the zero-one distribution. Thus minimizing these uncertainty measures is an intuitive and efficient approach to active learning, since minimizing these uncertainties also minimizes the entropy of the probability model. Recently, complex models were used in active learning [32, 28] focusing on labeling pixels, using standard approaches to evaluate the loss of a single pixel label. Our approach considers image patches thus requires to reason about exponentially many structured-labels within each patch. For this purpose we introduce new upper bounds to the entropy function to efficiently reason about images patches. In addition, we consider elaborate cost functions for annotating the boundaries of these image patches in multi-scale selection, thus providing a coarse-to-fine interactive annotation approach.

3 Active annotation framework

Image annotation is a costly procedure that currently prohibits from constructing large data sets that are crucial for developing better machine learning approaches. This is especially true for complex label spaces such as pixel labels for images, or location of joint positions in images of people. Often the structure over the label space can be compactly described using a probability distribution that captures the correlations between the labels. For pixels labels

these can be described using Markov random fields (MRFs) that enforce consistency between adjacent labels. For joint positions these priors can be derived from the kinematic structure of the underlying skeleton.

Let $y = (y_1, ..., y_n)$ be a set of structured labels that we wish to obtain for a given instance x. The joint probability distribution over the labels y is $p(y) \propto \exp(\theta(y; x, A_t))$. Here $A_t = \{a_1, ..., a_t\}$ is the set of annotations obtained till time t. Bayesian experimental design is a natural framework for actively seeking user inputs. Given a function $U(A)$ that measures the uncertainty of the labels given previous annotations, and a function $C(a)$ that measures the cost of an annotation a, Bayesian experimental design seeks to pick the annotation a that has the highest utility, i.e., uncertainty decrease per unit cost

$$a_t = \arg\max_{a \in \mathcal{A}} \frac{U(A_{t-1}) - U(A_{t-1} \cup a)}{C(a)} \tag{1}$$

The uncertainty $U(A_t) = H(p)$, is defined as the entropy, a commonly used uncertainty measure, over the label space

$$H(p) = -\sum_y p(y) \log p(y) \tag{2}$$

Thus, the Bayesian active annotation approach jointly minimizes the cost of annotations while reducing the uncertainty in the probability model. The main bottleneck in picking optimal annotations is the computation of the entropy, which is provably hard for general distributions. Unlike the entropy function, obtaining the best scoring annotation that is usually referred as the maximum a-posteriori (MAP) prediction, is often easier:

$$(MAP) \qquad \arg\max_{y_1,...,y_n} \theta(y_1, ..., y_n; x, A_t)$$

A notable example is the graph-cuts algorithm for binary pixel labeling problems in images. Significant effort has been spent on developing efficient solvers for commonly occurring problems. Motivated by this we propose a new way of computing entropy using MAP perturbations.

4 MAP perturbations

The quantities we are interested in can be computed whenever we can sample y according from the Gibbs distribution, $p(y) \propto \exp(\theta(y))$. We use $\theta(y)$ to refer to $\theta(y; x, A)$ in the interest of brevity. In high dimensional complex models sampling from the Gibbs distribution is provably hard whenever the model considers nonzero local evidence [22, 12]. Successful machine learning approaches strongly rely on informative local potentials, as they incorporate the data signal, e.g., image information. In this setting the resulting Gibbs probability landscape is often "ragged"; in such landscapes Markov chain Monte Carlo (MCMC) approaches to sampling from the Gibbs distribution may become prohibitively expensive. This is in contrast to the success of MCMC approaches in other settings (e.g., [23, 19]) where no data term (signal) exists. As an alternative, we suggest to use perturb-max models as posterior distributions [31]. The perturb-max models rely on random functions $\gamma_i : Y_i \to R$ for every pixel i. These random functions associate a random variable $\gamma_i(y_i)$ for each $y_i \in Y_i$. Thus these random perturbations may be used to measure the amount of change in the boundary of MAP decision:

$$p(\hat{y}) = P_\gamma\Big(\hat{y} = \arg\max_y \big\{\theta(y) + \sum_{i=1}^n \gamma_i(y_i)\big\}\Big) \tag{3}$$

Computing the entropy function (Equation (2)) is prohibitively expensive as it requires to sum over all possible labels, which are exponential in n. Here we propose to use an uncertainty measure that upper bounds the entropy and can be computed efficiently using MAP solvers. These upper bounds allow us to use Bayesian approaches for active learning efficiently even for exponentially large space of annotations. Our suggested entropy bounds measure the expected amount of change that is required in order to change the MAP decision.

Theorem 1 *Let $p(y)$ be a perturb-max probability distribution and let $\{\gamma_i(y_i)\}$ be a collection of i.i.d. random variables, each following the Gumbel distribution with zero mean, i.e., $P(\gamma_i(y_i) \le t) = \exp(-\exp(-(t + c)))$, where $c \approx 0.5772$ is the Euler-Mascheroni constant. Let y^*_γ be the perturbed MAP solution with respect to $\theta(y)$, i.e.,*

$$y^*_\gamma = \arg\max_y \big\{\theta(y) + \sum_{i=1}^n \gamma_i(y_i)\big\}$$

For simplifying the notation we denote y^*_γ by y^* while implicitly referring to its dependence on $\{\gamma_i(y_i)\}$. Then the following entropy bound holds:

$$H(p) \le E_\gamma\big[\sum_{i=1}^n \gamma_i(y^*_i)\big]$$

Proof: $\log Z(\theta) = \log\big(\sum_y \exp(\theta(y))\big)$ is the conjugate dual of the (minus) entropy function.

$$H(p) \quad = \quad \min_\theta \big\{\log Z(\theta) - \sum_y p(y)\theta(y)\big\}$$

Set $W(\theta) = E_\gamma\big[\max_y\{\theta(y) + \sum_{i=1}^n \gamma_i(y_i)\big]$, then $\log Z(\theta) \le W(\theta)$ whenever we consider the Gumbel perturbations, as shown in [14]. Therefore

$$H(p) \quad \le \quad \min_\theta \big\{W(\theta) - \sum_y p(y)\theta(y)\big\}$$

Finally, for perturb-max models

$$p(\hat{y}) = P_\gamma\Big(\hat{y} = \arg\max_y \big\{\hat{\theta}(y) + \sum_{i=1}^n \gamma_i(y_i)\big\}\Big)$$

there holds

$$\min_\theta \big\{W(\theta) - \sum_y p(y)\theta(y)\big\} = W(\hat{\theta}) - \sum_y p(y)\hat{\theta}(y).$$

Recalling that $W(\hat{\theta}) = E_\gamma\big[\max_y\{\hat{\theta}(y) + \sum_{i=1}^n \gamma_i(y_i)\big]$ we deduce $W(\hat{\theta}) = \sum_y p(y)\hat{\theta}(y) + E_\gamma\big[\sum_{i=1}^n \gamma_i(y_i^*)\big]$ and conclude the result. □

This entropy bound motivates the use of perturb-max posterior models. The computation of this bound relies on MAP solvers, thus it is significantly faster than the computation of the entropy itself whose complexity is generally exponential in n.

Using the linearity of expectation we may alternate summation and expectation. Thus the above theorem bounds the entropy by summing the expected change of MAP perturbations, $H(p) \leq \sum_i E_\gamma[\gamma_i(y_i^*)]$. This bound resembles to the independence bound for the entropy $H(p) \leq \sum_i H(p_i)$, where $p_i(y_i) = \sum_{y \setminus y_i} p(y)$ are the marginal probabilities [7]. The advantage of MAP perturbation bound over previous entropy bounds is that it only computes MAP assignments, while the standard entropy bounds require to compute the marginal probabilities, thus may be unpractical in high dimensional complex models. The independence bound is tight whenever the joint probability $p(y)$ is composed of independent systems, i.e., $p(y) = \prod_i p_i(y_i)$. interestingly, also the bound in Theorem 1 is tight in this setting.

Corollary 1 *Consider the setting in Theorem 1 and the independent probability distribution $p(y) = \prod_i p_i(y_i)$. Then*

$$H(p) = E_\gamma\big[\sum_i \gamma_i(y_i^*)\big]$$

Proof: Set $\theta_i(y_i) = \log p(y_i)$. Since $\log Z(\theta_i) = 0$ it follows from [14] that $E_{\gamma_i}\big[\max_{x_i}\{\theta_i(y_i) + \gamma_i(y_i)\}\big] = 0$. Since we are using i.i.d. Gumbel random variables

$$E_{\gamma_i}\big[\max_{y_i}\{\theta_i(y_i) + \gamma_i(y_i)\}\big] = \sum_{y_i} p(y_i)\theta_i(y_i) + E_{\gamma_i}\big[\gamma_i(y_i^*)\big]$$

while the left hand side equals zero, and the right hand side contains the quantities of interest. We conclude while setting $\theta(y) = \sum_i \theta_i(y_i)$. □

There are two special cases for independent systems. First, the zero-one probability model, for which $p(y) = 0$ except for a single configuration $p(\hat{y}) = 1$. The entropy of such probability distribution is minimal, i.e., zero, as they have no uncertainty. In this case, the MAP perturbation entropy bound assigns $y_\gamma^* = \hat{y}$ for all random functions γ_i. Since these random variables have zero mean, it follows that $E_\gamma\big[\sum_i \gamma_i(\hat{y}_i)\big] = 0$. Another important case is for the uniform distribution, $p(y) = 1/|Y|$ for every $y \in Y$. The entropy of such a probability distribution is maximal, i.e., $\log|Y|$, as it has maximal uncertainty. To estimate our entropy bound for the uniform distribution we first note that since $Y = Y_1 \times \cdots Y_n$ is a discrete product space $\log|Y| = \sum_i \log|Y_i|$. Also, the uniform distribution is coupled with $\theta(y) \equiv 0$ thus the MAP perturbation entropy bound equals to $\sum_i E_{\gamma_i}\big[\max_{y_i} \gamma_i(y_i)\big]$. Since the random perturbations follow the Gumbel distribution, this entropy bound is tight, i.e., $E_{\gamma_i}\big[\max_{y_i} \gamma_i(y_i)\big] = \log|Y_i|$ (cf. [14]). This suggests that the MAP perturbations can be used as an alternative uncertainty measure:

Corollary 2 *Consider the setting in Theorem 1. Set*

$$U(p) = E_\gamma\big[\sum_i \gamma_i(y_i^*)\big] \tag{4}$$

Then $U(p)$ is an uncertainty measure, i.e., it is nonnegative, it attains its minimal value for the zero-one distribution and its maximal value for the uniform distribution.

Proof: It is nonnegative since it upper bounds the entropy function. $U(p)$ attain its minimal value when $p(y)$ is the zero-one distribution since $U(p) = 0$ in this case (see discussion above). Lastly, we prove that $U(p)$ attains its maximal value for the uniform distribution, or equivalently when $\theta(y) \equiv c$ for any constant c. Assume the contrary, thus there are $y, \hat{y}$ for which $\theta(y) < \theta(\hat{y})$. Thus there are $\sum_i \gamma_i(y) > \sum_i \gamma_i(\hat{y}_i)$ although $y^* = \hat{y}$, a contradiction. □

The advantage of using the MAP perturbations uncertainty measure over standard Bayesian active learning approaches is that it does not require MCMC sampling procedures. Therefore, our approach well fits high dimensional complex models that currently dominate machine learning applications such as computer vision. It addition, the MAP perturbations uncertainty measure upper bounds the entropy thus reducing its uncertainty effectively reduces the entropy.

5 Active boundary annotation

The ability to sample and reason about uncertainties provides a powerful framework for active annotations of structured labels. In particular complex annotation tasks with varying costs can be considered. To illustrate this, we consider the task of obtaining pixel accurate boundaries of objects in images. In typical high-resolution images labeling a accurate boundary can take an order of magnitude more time than a coarse one. In our approach we start with a coarse boundary and actively refine it. As seen in Fig. 1, at each step the user is shown an region of the image α and

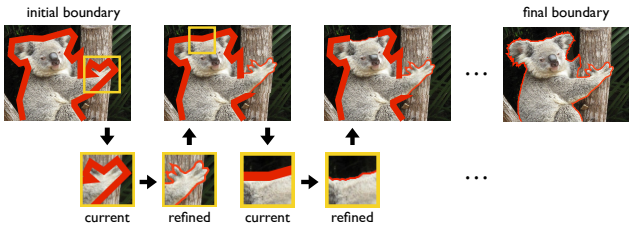

Figure 1: We actively suggest regions for boundary refinement based on label uncertainty and annotation cost. Naively annotating the boundary on the right requires 20× more points than the initial one.

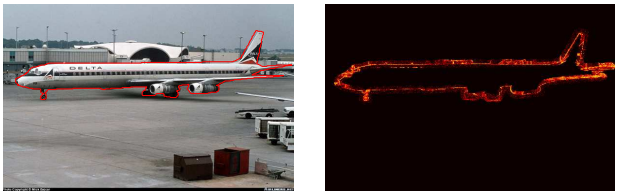

Image + initial boundary Per-pixel entropy bound

Figure 2: Pixel-wise entropy bound estimated on the image using 100 samples from perturb-max model visualized as a heat map (red is high, black is low). The model is uncertain in the regions near the wings, nose, and the wheels.

is asked to refine the boundary represented as a piecewise-linear polygon by moving the points on the boundary. Once the annotation a for the region α is obtained, it is incorporated into the model to obtain the new probability distribution over pixel labels.

In contrast to previous approaches for active labeling where the annotation task is to mark a pixel labels [32], our task is more intuitive and provides labels for many pixels simultaneously. Labelling "super pixels" is another approach, but suffers if the initial segmentation misses the true boundaries. For objects with thin structures this can be a challenge. Brush strokes, and other input modalities present a challenge on how to model the space of user input. See [29] for a discussion on interactive segmentation systems. Boundaries on the other hand can be compactly described by a starting and end point on the curve and are naturally multi-scale. This allows us to reason over the space of possible annotations in a coarse-to-fine manner while considering complex cost functions that depend on the boundary complexity.

From a labeling of a subset of pixels one can obtain a full labeling using the "grabcut" model [3, 2]. In this setting a possible annotation for an image x with n pixels is describe by the $n-$tuple $y = (y_1, ..., y_n)$. Each pixel label is either foreground or background, namely, $y_i \in \{-1, 1\}$. The local potential functions $\theta_i(y_i; x)$ are derived from Gaussian mixtures on the pixel color values to model $\log P(y_i|x)$ that provide an initial foreground/background preference for every pixel. The pairwise potential functions $\theta_{i,j}(y_i, y_j; x) = \exp(-(x_i - x_j)^2)y_i y_j$, where x_i denotes the intensity of image x at pixel i, encourage adjacent pixels with the same color to share the same labels. The quality of each labeling is described by the global potential functions

$$\theta(y_1, ..., y_n; x) = \sum_{i=1}^{n} \theta_i(y_i; x) + \sum_{(i,j)\in E} \theta_{i,j}(y_i, y_j; x) \quad (5)$$

The best scoring labeling for a given image is the maximum a-posteriori (MAP) prediction, which in our setting my be derived using the graph-cuts algorithm. The MAP prediction approach is computationally appealing. Although the total number of possible annotations is exponential in n, the graph-cuts algorithm recovers the MAP prediction in a linear time.

The uncertainty in labels after annotating an image patch α at iteration t is the expected uncertainty of the labels given the annotation according to the "user model" P_u, i.e.,

$$U_t(\alpha) = \mathrm{E}_{a\sim P_u(a;\alpha)} U(A_{t-1} \cup a) \quad (6)$$

where, $U(A_{t-1} \cup a)$ is defined in Eqn. 2. Fig. 2 shows the uncertainty (entropy) computed using our method given the image and the initial boundary visualized as a heat map.

Similarly the cost of annotating a patch α is the expected cost of annotating α according to the user model, i.e.

$$C(\alpha) = \mathrm{E}_{a\sim P_u(a;\alpha)} C(a) \quad (7)$$

where, $C(a)$ measures the cost of labeling a segmentation. Since our annotation task is boundary marking, the cost is more appropriately measured as the boundary complexity of the segmentation. This is estimated by approximating the boundary of the segmentation with a piecewise-linear curve that is within τ (the desired precision level) of the original boundary. The cost is the number of points in the polygon approximation.

In practice since we don't have access to the user model we approximate this using the model $\theta(y)$. Concretely we first sample $y \sim \theta(y)$ using the perturb-max framework, and restrict the labels to the patch α. This is then resized to the desired display resolution. Note that α may be of any size, thus it allows us to annotate image patches in any scale.

Intuitively, our Bayesian active learning approach for boundary annotation suggests the human expert an image patch that is easy (or cheap) to annotate, yet the algorithm is uncertain about its correct annotation. Fig. 3 shows an example of how the quantities such as cost and certainties are computed using the perturb-max framework. The sampled segmentations approximate what a user might have annotated. From these samples the average boundary annotation cost and certainties are computed.

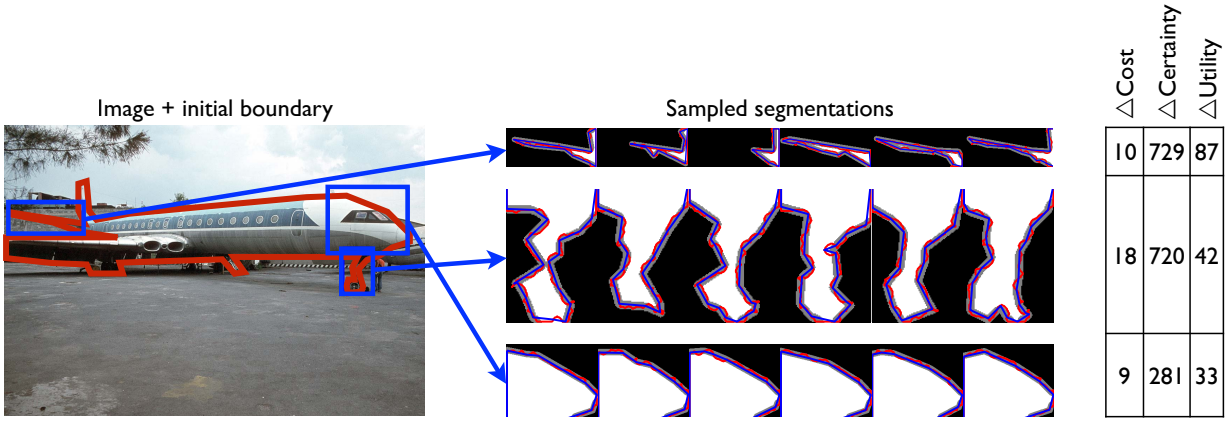

Figure 3: Uncertainty and cost tradeoffs computed for three different regions α in the image shown in blue in the leftmost image. In the middle are samples generated according to $p(y_\alpha|x, A)$, shown along with the inferred user annotated boundary (in blue), exact boundary (in red), as well as the figure, ground, and unknown regions as white, black, and gray pixels respectively. While the region near the front wheel offers higher reductions in uncertainty, it is expensive to annotate. On the other hand, the region near the nose is cheap to annotate, but offers lower reductions in uncertainty. The region near the wing offers the best tradeoff between uncertainty and cost among the three, and will be picked by the active learner.

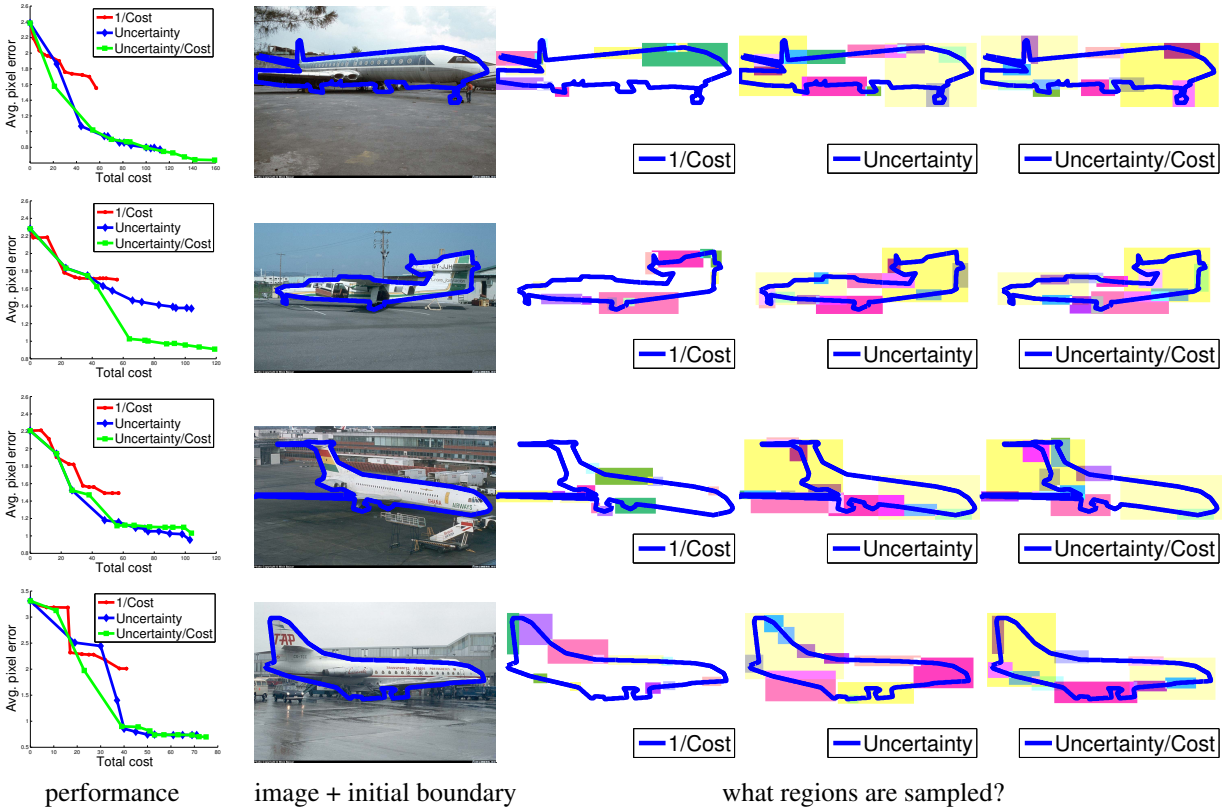

Figure 4: (Left) Performance of various active learning strategies. (Middle) Image and the initial boundary. (Right) The regions sampled by active learning strategies for refinement in the first 10 iterations of the algorithm based on cost, uncertainty, and utility=uncertainty/cost.

5.1 Experimental evaluation

We present experiments on a dataset of airplane images[1] for which we have obtained pixel accurate annotations. Man-made objects like airplanes have complex structures and obtaining pixel accurate boundaries is painstakingly slow – it took about 15 minutes on average to annotate each image of the 10 images. In stark contrast, obtaining approximate boundaries takes less than 30 seconds.

We present an experimental study where we assume that the annotator provides a piecewise-linear approximation of the ground-truth boundary for the specified region within a specified tolerance $\tau = 1\%$ of the size of the display window. This closely approximates our annotation tool where the annotators can mark the boundaries as polygons. At any step in the annotation process we can measure the *error* as total number of incorrectly assigned pixels with respect to the accurate ground-truth segmentation *divided* by the length of the ground-truth boundary. We can also measure the *cost* as the total number of points on the boundary that have been annotated so far. Since we have access to the ground truth, we can compute these quantities automatically and report performance of various techniques in terms of 'pixel error vs. cost' tradeoff curves.

We consider two baselines, oracle that achieves error equal to the ground truth boundary at cost n, where n is the number of points in the piecewise-linear representation of the boundary of the exact annotation, and agnostic which achieves zero error at cost m, where m is the number of points in the boundary of the annotation at the resolution of the image. Typically we have $m >> n$. The oracle is equivalent to annotator who can automatically zoom in to the right regions and annotate boundaries based on the complexity of the boundary, whereas agnostic simply traces the boundary at the pixel level and provides a upper bound on the cost. This would not be possible without a very high resolution display. Note that the oracle does not use bottom-up image information and the user has to trace boundaries along high-curvature regions to obtain precise boundaries, even though they may have high figure-ground contrast, thereby wasting effort.

We consider three active learning strategies that try to minimize the pixel error. First strategy called active-certain picks regions that provide the reduction in uncertainty, the second picks regions that are cheapest to annotate active-cost, and the last strategy called active-util picks regions that provide the highest reduction in uncertainty per unit cost (Equation (1)).

For each image, typically of size 10^5-10^6 pixels, we start with an initial approximate boundary and actively suggest regions of maximum dimensions between 50-250 pixels for refinement. There are a very large number of possible regions in the image to consider, so we approximate the scheme by randomly sampling 10 non-overlapping regions from the image and pick the best one (depending on the criteria) at each time step. For speed we update the unary potentials only when there is a significant change in the number of pixels that have changed their labels.

[1] downloaded from http://airliners.net

5.2 Active annotation results

Fig. 5(a) shows the results of active annotation for various methods averaged over the images dataset. From the figure we can see that the active learning strategies active-util/certain outperform the active-cost. Fig. 4 shows the regions sampled by different active learning strategies on several images.The strategy based on cost focusses on the linear segments as they are inexpensive to annotate but may not be uncertain to benefit from refinement. On the other hand, methods based on the certainty and utility refines regions with high errors and achieve faster error reduction.

The active learning strategies outperform the oracle for obtaining pixel-accurate boundary, i.e., pixel error = 1, showing the advantage of using bottom-up image cues. The active learning strategies based on utility and certainty require about 50 points on average compared to the oracle that requires 75 points (Fig. 5(a)). Moreover, the active strategy makes for an intuitive annotation tool as as it automatically suggests regions to refine instead of the user having to manually zoom in. A preliminary user study suggests that this can save up to $4\times$ in annotation time. Moreover, non-overlapping regions can be refined in parallel making it more suitable for crowdsourcing.

How well do the predicted uncertainty and cost match the truth? Fig. 5(c)-(d) show scatterplots of the expected and actual costs, as well as those for errors. The error is predicted quite well, but the cost of annotation is typically over-estimated. This is because MAP estimation tends to produce segmentations with jagged edges. Better cost estimates may be obtained by conducting user studies.

Does multi-scale annotation help? Our active learning strategy considers regions across multiple scales, adaptively zooming in on regions based on the uncertainty and cost tradeoff they offer. As a comparison we also compute results by restricting the sizes of the selected to a smaller range of a maximum width of 75-125 pixels, compared to 50-250 pixels. Fig. 5(b) shows that the multi-scale strategy can provides a saving of about 20% in cost.

6 Conclusion and future work

High quality training data sets are important to develop novel approaches that can solve challenging tasks. Unfortunately, attaining high quality annotations for complex models is an expensive and time consuming task as it in-

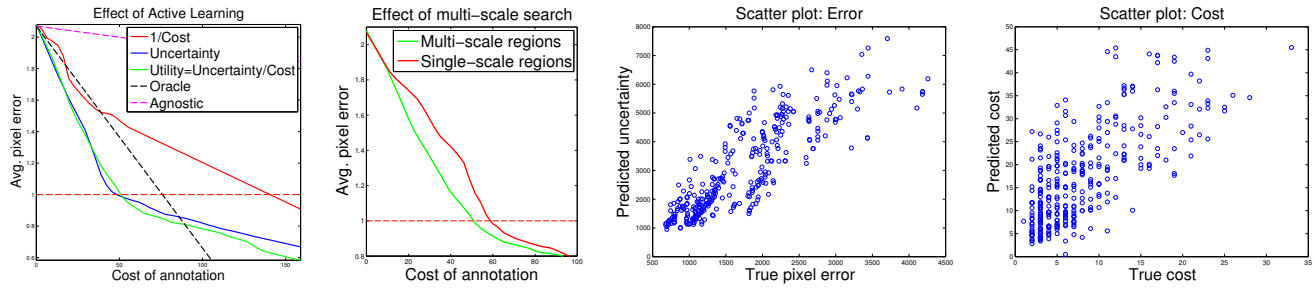

Figure 5: (a) *Error vs. cost tradeoffs for various annotation methods*. Using active learning we are able to annotate the boundaries within an avg. error of 1 pixel at about 66% of the cost required for annotating using the oracle. (b) *Effect of multi-scale.* Multi-scale selection of regions provide a saving of about 20% in cost over single-scale regions. (c,d) *Quality of error and cost estimation.* Scatter plot of predicted and true error (c), and same for the cost (d).

volves an expert annotator to process each instance. In this work we use Bayesian active learning to suggest simpler annotation tasks sequentially in order to reduce the cost and show its application for precise boundary labeling.

To avoid the expensive MCMC sampling techniques we propose a new uncertainty measure that is based on MAP perturbations: by randomly perturbing the boundary of decision, the algorithm is able to estimate its uncertainty. This results in a novel upper bound for the entropy, which is different than the standard entropy bounds that are defined over marginal probabilities. Since entropy has influenced research in many fields, this entropy bound might be of independent research beyond the scope of active learning.

Our approach uses MAP solver as a building block to compute uncertainty, thus we are able to apply Bayesian active learning in order to achieve high-quality boundary annotations of images. We show that our approach significantly reduce the cost and the run-time of expert annotators. The results here can be taken in several different directions. Theoretically, it would be interesting to provide MAP perturbations entropy bounds with higher dimensional perturbations. These bounds might compensate overlapping perturbations by their covering number, e.g., [17].

Finally, our approach of Bayesian active learning is applicable whenever MAP perturbation can be solved efficiently. We aim to explore problems beyond super-modular potential functions in the context of active learning. For example reasoning over matchings using our MAP perturbations to annotate machine translation data sets faster.

References

[1] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009. 2

[2] Andrew Blake, Carsten Rother, Matthew Brown, Patrick Perez, and Philip Torr. Interactive image segmentation using an adaptive gmmrf model. In *ECCV*. 2004. 5

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001. 1, 5

[4] Steve Branson, Pietro Perona, and Serge Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *ICCV*, 2011. 2

[5] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995. 2

[6] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *arXiv preprint cs/9603104*, 1996. 2

[7] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 4

[8] Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *Learning Theory*, pages 249–263. Springer, 2005. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 2

[10] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, 1997. 1, 2

[11] A. Gane, T. Hazan, and T. Jaakkola. Learning with maximum a-posteriori perturbation models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014. 1, 2

[12] L.A. Goldberg and M. Jerrum. The complexity of ferromagnetic ising with local fields. *Combinatorics Probability and Computing*, 16(1):43, 2007. 2, 3

[13] Steve Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007. 2

[14] T. Hazan and T. Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *ICML*, 2012. 1, 2, 3, 4

[15] T. Hazan, S. Maji, Keshet J., and T. Jaakkola. Learning efficient random maximum a-posteriori predictors with non-decomposable loss functions. *Advances in Neural Information Processing Systems*, 2013. 1, 2

[16] T. Hazan, S. Maji, and T. Jaakkola. On sampling from the gibbs distribution with random maximum a-posteriori perturbations. *Advances in Neural Information Processing Systems*, 2013. 1, 2

[17] T. Hazan, J. Peng, and A. Shashua. Tightening fractional covering upper bounds on the partition function for high-order region graphs. 2012. 8

[18] Neil Houlsby, Jose Miguel Hernandez-Lobato, Ferenc Huszar, and Zoubin Ghahramani. Collaborative gaussian processes for preference learning. In *NIPS*, 2012. 1, 2

[19] Mark Huber. A bounding chain for swendsen-wang. *Random Structures & Algorithms*, 22(1):43–59, 2003. 3

[20] Tomoharu Iwata, Neil Houlsby, and Zoubin Ghahramani. Active learning for interactive visualization. In *AISTATS*, 2009. 2

[21] Tommi Jaakkola and Hava Siegelmann. Active information retrieval. 2001. 2

[22] M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993. 3

[23] M. Jerrum, A. Sinclair, and E. Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004. 3

[24] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, 2007. 1, 2

[25] J. Keshet, D. McAllester, and T. Hazan. Pac-bayesian approach for minimization of phoneme error rate. In *ICASSP*, 2011. 1, 2

[26] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 28(10), 2006. 1

[27] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008. 2

[28] Wenjie Luo, Alex Schwing, and Raquel Urtasun. Latent structured active learning. In *Advances in Neural Information Processing Systems*, pages 728–736, 2013. 2

[29] Hannes Nickisch, Carsten Rother, Pushmeet Kohli, and Christoph Rhemann. Learning an interactive segmentation system. In *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing*, 2010. 5

[30] Francesco Orabona, Tamir Hazan, Anand D Sarwate, and Tommi. Jaakkola. On measure concentration of random maximum a-posteriori perturbations. *arXiv:1310.4227*, 2013. 1, 2

[31] G. Papandreou and A. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*, Barcelona, Spain, November 2011. 1, 2, 3

[32] Gemma Roig, Xavier Boix, Roderick de Nijs, and Sebastian Ramos. Active map inference in crfs for efficient semantic segmentation. In *ICCV*, 2013. 1, 2, 5

[33] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Machine Learning: ECML 2006*, pages 413–424. Springer, 2006. 2

[34] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008. 2

[35] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010. 1

[36] Pannaga Shivaswamy and Thorsten Joachims. Online structured prediction via coactive learning. *arXiv preprint arXiv:1205.4213*, 2012. 1, 2

[37] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In *UAI*, 2008. 1

[38] D. Tarlow, R.P. Adams, and R.S. Zemel. Randomized optimum models for structured prediction. In *AISTATS*, 2012. 1, 2

[39] Simon Tong and Daphne Koller. Active learning for parameter estimation in bayesian networks. In *NIPS*, 2000. 2

[40] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002. 2

[41] L.G. Valiant. The complexity of computing the permanent. *Theoretical computer science*, 8(2):189–201, 1979. 1

[42] Sudheendra Vijayanarasimhan and Kristen Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *CVPR*, 2009. 2

[43] Sudheendra Vijayanarasimhan and Kristen Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, 2011. 1, 2

[44] Carl Vondrick and Deva Ramanan. Video annotation and tracking with active learning. In *NIPS*, 2011. 1, 2

[45] Peter Welinder and Pietro Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *CVPR*, 2010. 2

[46] J. Zhang, H. Liang, and F. Bai. Approximating partition functions of the two-state spin system. *Information Processing Letters*, 111(14):702–710, 2011. 2