Active Learning for Semantic Segmentation with Expected Change

Alexander Vezhnevets[1] Joachim M. Buhmann[1] Vittorio Ferrari[2]
[1]ETH Zurich Zurich, Switzerland
[2]The University of Edinburgh Edinburgh, UK

Abstract

We address the problem of semantic segmentation: classifying each pixel in an image according to the semantic class it belongs to (e.g. dog, road, car). Most existing methods train from fully supervised images, where each pixel is annotated by a class label. To reduce the annotation effort, recently a few weakly supervised approaches emerged. These require only image labels indicating which classes are present. Although their performance reaches a satisfactory level, there is still a substantial gap between the accuracy of fully and weakly supervised methods. We address this gap with a novel active learning method specifically suited for this setting. We model the problem as a pairwise CRF and cast active learning as finding its most informative nodes. These nodes induce the largest expected change in the overall CRF state, after revealing their true label. Our criterion is equivalent to maximizing an upper-bound on accuracy gain. Experiments on two data-sets show that our method achieves 97% *percent of the accuracy of the corresponding fully supervised model, while querying less than* 17% *of the (super-)pixel labels.*

1. Introduction

In this paper, we consider the problem of *semantic segmentation*, where a label must be predicted for every pixel in an image (e.g. "dog", "car" or "road"). Semantic segmentation has recently attracted a lot of attention [1, 2, 3, 4, 5]. The standard approach is to train with full supervision, where every pixel is manually labeled by a human annotator. Producing this annotation is very time-consuming. Recently, a few *weakly supervised* methods have emerged [6, 7, 8], which can train from image labels indicating which classes are present, but their location is unknown. Although weakly supervised methods reach a good performance, there is still a significant gap between weakly and fully supervised methods. In this paper, we try to bridge this gap by *active learning* (fig. 1).

As in most existing works [1, 2, 3, 5, 8, 6], we model the problem with a pairwise conditional random field (CRF), which we define over superpixels. The unary potential carry appearance models to classify a superpixel based on image measurements, while the pairwise potential encourages connected superpixels to assume the same label. In our setting, the training images are initially weakly labeled. The label of each superpixel they contain is a latent variable in the CRF. First, we train a weakly supervised model Multi Image Model (MIM) [6] to recover a first approximation of these labels. Then we run a active learning algorithm which queries the oracle for the true state of a few latent variables selected by a novel criterion. When the true state of a variable is revealed, it *induces change to the state of other variables as well*. These changes propagate locally through the pairwise potentials of the CRF, as well as globally through the unary potential, because the appearance models are retrained according to the newly revealed label. Due to this long-range interaction, changes can reach very far, often propagating to several other images.

There are relatively few works on active learning in computer vision [9, 10, 11, 12]. One criterion for choosing queries is *uncertainty sampling* [11, 12]. However, this criterion can be misguided in two ways. First, after the oracle reveals the state of an uncertain variable, it may induce changes in the state of only a few other variables, thus having a low impact. Second, a valuable query might be missed due to a false certainty about a variable label. In this paper, we propose a method based on a different criterion: query for the labels of variables that induce the largest expected change (EC) in the labeling of the whole training set. We show this method to directly maximizes the expectation of an upper-bound in accuracy improvement over the training set. Computing our score naively would be prohibitively expensive, as it involves retraining appearance model parameters and rerunning CRF inference, for each latent variable (superpixel) and each possible label. We show how to employ dynamic graph cuts [13] to greatly reduce computational cost. While computing the EC score we also estimate the *influence area* of each latent variable, i.e. the subset of variables which are expected to change if its true label is revealed. This information suggests a further speedup by querying for more than a single variable per active learning cycle. To maximize efficiency, in terms of how much

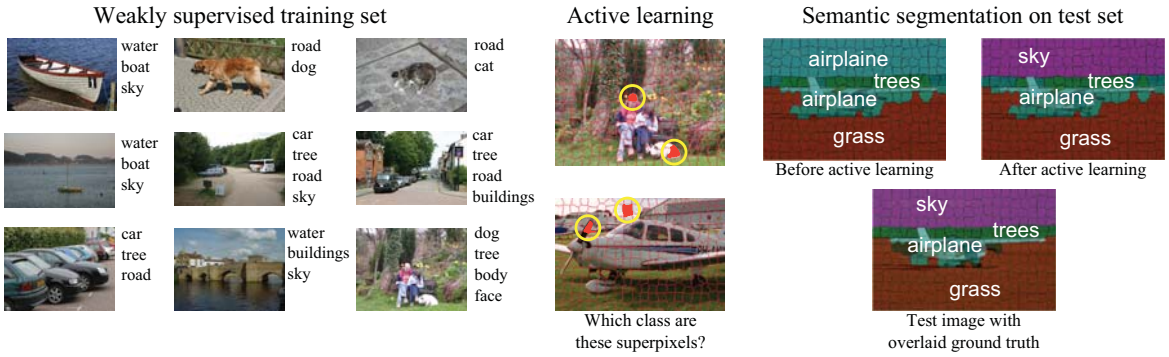


Figure 1. **weakly supervised semantic segmentation with active learning.** The input is a weakly supervised training set, where images are labeled by the classes they contain. The active learning proceeds by querying an oracle for the true label of superpixels selected by a specialized criterion. We use acquired information to classify superpixels in previously unseen test images.

change is achievable from a fixed number of queries, we choose the variables with the highest EC scores whose influence areas do not overlap. The combined effect of the two speedups renders our change-driven strategy computationally feasible. Note that this change-driven strategy does not suffer from the problems of uncertainty sampling, as it maximizes the expected impact of the queries, measured by the total change in the CRF state. A related criterion was proposed before [9, 10], but never for structured models like CRFs.

Experiments on two popular benchmark data-sets MSRC-21 and the subset of LabelMe used by [14] demonstrate that our method outperforms uncertainty sampling for semantic segmentation. Our approach achieves 97% percent of the accuracy of the corresponding fully supervised model, while querying less than 17% of the (super-)pixel labels. This provides an insight into the latent structure of semantic segmentation data, i.e. there are strong and far-reaching dependencies between labels of different superpixels, that span over several images. Because of these relations, knowledge of even a small subset of the labels allows us to determine most of the other labels.

2. Related work

Semantic segmentation is represented by many fully supervised methods [1, 2, 3, 4, 5] and a few weakly supervised ones [6, 7, 8]. In computer vision, active learning was mostly used for the task of object detection or whole-image classification [15, 11, 12, 16]. To our knowledge, active learning for semantic segmentation was considered by only very few works [9, 10]. Vijayanarasimhan et al. [9] focus on predicting the cost of labeling (in seconds of annotators time) and trading it for informativeness of the query. Siddiquie et al. [10] introduce contextual queries, such as "*is sky above ground?*" and model them as special edges in a pairwise CRF. Analog to us, [9, 15] compute the reduction in expected misclassication risk over all of the training data in order to assess the influence of a potential query. Unlike these works, our model is *structured* - it includes pairwise connections between variables (superpixels).

The problem of selecting the most informative subset of variables in a graphical model was studied by [17, 18, 19]. Krause et al. [17] present optimal algorithms for computing and optimizing the value of information on a chain graphical models. In another paper [18] they address the same problem for Bayesian networks. In [19] a fast way to evaluate the informativeness of a variable in a graphical model was proposed. All of these works are concerned with setting where exact inference is possible (trees, Bayesian networks). CRF models for the semantic segmentation task have a rough grid structure where nodes have a state space of about 20-50 labels, thus exact inference is not feasible. To our knowledge, we are the first to go beyond uncertainty sampling for such CRF models.

Another relevant work [20] considers the task of active multiple instance learning. It is essentially a two-class version of our setting. The method builds upon multiple instance logistic regression. They introduce the expected gradient length (EGL) criterion: choose the query which maximizes the expectation of the gradient for the regression coefficients. This heuristic is similar in spirit to our expected change criterion. However, we quantify change differently, as our model (multilabel CRF) is structured and a gradient is not well-defined for it. We also show that our strategy directly maximizes the expected upper-bound on accuracy improvement, which is not true for EGL.

3. weakly supervised semantic segmentation with a pairwise CRF

We follow the method introduced in [6], for self-consistency and to introduce notation we briefly review it below. Images are represented by their superpixels, obtained by an oversegmentation algorithm [21]. Let $\tau =$

$\left\{I^j = \left(\{x_i^j\}_{i=1}^{N_j}, Y^j\right)\right\}_{j=1}^{N}$ be the training set, where image I^j consists of superpixels x_i^j. For each image, we are given a label set $Y^j \subset \mathcal{Y}$, which is a subset of the set of all possible labels $\mathcal{Y} = \{1, ..., C\}$, corresponding to classes. Each superpixel x_i^j has an associated *latent* label $y_i^j \in Y^j$. The image label set Y^j is the union over the (unknown) labels of all superpixels in the image ($Y^j = \bigcup y_i^j$). The task of weakly supervised learning is to recover the latent labels y_i^j and to learn appearance models for the classes. These will later help to predict superpixel labels in new test images.

Model. We model the weakly labeled training set as a CRF, where nodes correspond to the latent superpixel labels. The total energy $\mathcal{E}$ of the model is a function of these labels y_i^j and appearance model parameters θ

$$\mathcal{E}\left(\{y_i^j\}, \theta\right) = \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) + \sum_{(y_i^j, y_{i'}^{j'}) \in E} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) \quad (1)$$

The first unary potential $\psi\left(y_i^j, x_i^j, \theta\right)$ measures how well the appearance of x_i^j matches the appearance model $\theta_{y_i^j}$ of class y_i^j. If $f(x, \theta) \rightarrow \mathcal{R}^C$ is a multiclass classifier outputting probabilities $f_y(x, \theta)$ for superpixel x taking label y, then we can define the unary potential as $\psi(y, x, \theta) = -\log f_y(x, \theta)$. In this paper we consider a Naive Bayes as an appearance model. The second unary potential $\pi(y_i^j, Y_i^j)$ makes sure that a superpixel can only take a label from the label set Y^j of the image

$$\pi(y_i^j, Y_i^j) = \begin{cases} \infty & y_i^j \notin Y^j \\ 0 & y_i^j \in Y^j \end{cases} \quad (2)$$

A CRF labeling which respects this constraint for all nodes is called *admissible*.

The pairwise potential ϕ encourages connected superpixels to take the same label if their appearance similarity is high

$$\phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) = \begin{cases} 1 - D\left(x_i^j, x_{i'}^{j'}\right) & y_i^j \neq y_{i'}^{j'} \\ 0 & y_i^j = y_{i'}^{j'} \end{cases} \quad (3)$$

where $D\left(x_i^j, x_{i'}^{j'}\right)$ is a similarity metric between two superpixels, scaled to $[0, 1]$. Our particular choice of similarity metric is discussed in sec. 5. Note how these potentials are submodular, since $1 - D\left(x_i^j, x_{i'}^{j'}\right) \geq 0$ always. The pairwise potentials are defined on the set of edges E. Usually, these edges connect spatial neighbours in the same image [8, 1, 2, 3, 5]. We show later that extending E to include edges between superpixels from different images, as originally done in [6] can significantly improve performance.

Weakly supervised learning. We minimize the energy in eq. (1) by alternating minimization. Starting from an admissible random labeling, we alternate between learning appearance model parameters θ and inferring the latent labels y_i^j. When keeping the labeling fixed and assuming our features are histograms, learning a Naive Bayes classifier parameters has a closed form solution. Let θ_l be the likelihood vector, such that $\theta_l^i = P(x_i|l)$; then $\theta_l = XY^l$, where X is a matrix of superpixel features and Y^l is a binary column-vector with $Y_i^l = 1$, when $y_i = l$. If Y is a matrix, who's columns are Y^l, then parameter matrix $\theta = XY$ can be obtained by just one matrix multiplication, plus one more multiplication for if we want to incorporate a prior. Our particular choice of features is described in sec. 5. When keeping the appearance parameters fixed, the energy (1) is submodular. Since it is a multilabel problem, we cannot obtain global minima, at least not exact, but we can efficiently find a good approximation using the alpha-expansion algorithm [22, 23, 24, 25]. Therefore, we alternate between these two steps to efficiently determine both θ and $\{y_i^j\}$.

CRF structure. Most approaches [1, 2, 3, 5, 8] establish connections E (eq. (1)) only between neighbouring superpixels in one image. This produces a set of disconnected components, sharing only the appearance models θ. As proposed in [6], we establish connections also between superpixels in different images. We do so by connecting superpixels from images that share a label: $Y^j \bigcap Y^{j'} \neq \emptyset$. For each superpixel y_i^j, we first select the $q = 3$ most similar superpixels from each other image sharing a label with Y^j. We then establish connections from y_i^j to the top $p = 21$ selected superpixels. The reason for this procedure is to add only a moderate number of most-valuable connections (to keep inference fast and memory requirements low). Recall that connections between superpixels with very different appearance have little influence, as eq. (3) is near 0 for any labeling. Appearance similarity is measured by D (defined above). As in [26], we can interpret E as a model for the manifolds formed by superpixels in the space defined by the similarity metric D. Pairwise potentials penalize labelling $\{y_i^j\}$ that cut through these manifolds.

4. Active Learning

The active learning stage starts from the output of weakly supervised learning, i.e. a partially incorrect labeling. During active learning, the computer can submit a query (i, j) to an oracle O, that reveals $O(i, j) = l_i^j$ the true state l_i^j of a latent variable y_i^j, i.e. the label of the corresponding superpixel. The goal is to achieve the maximum increase in accuracy over the whole training set τ with a minimal number of queries.

The active learning protocol that we consider is summarized in Alg. 1. First, all possible queries $\Omega = \{(i, j) | y_i^j \in$

$\tau\}$ are evaluated according to some score function $S : \Omega \rightarrow \mathcal{R}$ (e.g. uncertainty). Next, a query set $Q \subset \Omega$ consisting of one or more queries is selected by a rule U and is submitted to the oracle. Usually [11, 12, 16, 9, 10] U selects only one query, according to S, or the highest scored k queries. However, later we discuss other possible forms of U. After the oracle delivers the answers $O(Q) = \{l_i^j | (i,j) \in Q\}$, the model is retrained and the procedure restarts. To integrate the revealed state l_i^j of a variable y_i^j into the model, we set $y_i^j = l_i^j$.

Let $F(x, \theta)$ be the output (a label) of the CRF for a training superpixel x, with the appearance model parameter vector θ. Also, let $F(x, \theta^t | y_i^j = l)$, be the output of the model, given that the latent variable y_i^j is assigned to label l (implying that θ^t is relearnt accordingly). The Expected Change (EC) score of y_i^j is defined as

$$EC(i,j) = \frac{1}{|Y^j|} \sum_{l \in Y^j} \sum_{i' \neq i, j' \neq j} w_i^j \left[F\left(x_{i'}^{j'}, \theta^t | y_i^j = l\right) \neq F\left(x_{i'}^{j'}, \theta^t\right) \right] \tag{4}$$

Here w_i^j is the importance weight of y_i^j, i.e. the number of pixels in superpixel x_i^j. The EC score measures the expected amount of change in the CRF, measured as the weighted change in superpixel labels over the whole training set. It can be shown that querying the oracle for the y_i^j with the largest $EC(i,j)$ maximizes the expectation of the upper-bound of the accuracy improvement over the training set

$$\sum_{i,j} w_i^j \left[F(x_i^j, \theta^{t+1}) = l_i^j \right] - \sum_{i,j} w_i^j \left[F\left(x_i^j, \theta^t\right) = l_i^j \right] \leq \sum_{i,j} w_i^j \left[F\left(x_i^j, \theta^{t+1}\right) \neq F\left(x_i^j, \theta^t\right) \right] \tag{5}$$

where l_i^j is the true label of y_i^j. Since true labels are not known, we cannot maximize the bound directly and must take the expectation over all admissible labels instead (eq. (4)).

Computational cost. To evaluate the EC score for each latent variable, we have to run through all (still unknown) latent variables. For each admissible label of each variable we must do (Alg.2, fig. 2): i) retrain appearance models θ for two classes (the former label y_i^j and the hypothesized label l); ii) infer new labeling with alpha-expansion; iii) record the change. This amounts to NK retraining and inference runs, where K is an average number of admissible labels per variable and N is the total number of latent variables that are still unknown. Retraining θ is quite fast. It involves only three matrix multiplications, as we have to update only two appearance models. However, the alpha-expansion step over our large CRF is very slow. Therefore, in a naive implementation, the computational cost of the algorithm is enormous. In the following subsections, we propose two techniques to accelerate the algorithm. The first is based on recycling computations between graph-cut runs in the inference stage (step 1.a.iii of Alg. 2). The second technique is a new selection rule U that queries for several labels in a batch, while avoiding redundant queries (i.e. inducing changes to similar set of variables).

Algorithm 1 Generic active learning procedure

Input: Training set τ, initial parameters θ^0, initial labeling $L^0 = \{y_i^j = F(x_i^j, \theta^0)\}$, maximum number M of queries to the oracle O, query scoring function S, query selection rule U.
Output: updated labeling and parameters θ^*

1. $t = 0$ and $m = 0$
2. **while** $m < M$
   (a) **for each** unknown latent variable y_i^j, evaluate $S(i,j)$
   (b) Select query set Q with selection rule U
   (c) Query the oracle for the labels $l_i^j = O(i,j),\ \forall (i,j) \in Q$
   (d) Set $y_i^j = l_i^j\ \forall (i,j) \in Q$
   (e) Retrain appearance models θ^{t+1} and infer latent variable labels
   (f) $m = m + |Q|$ and $t = t + 1$
3. return $\theta^* = \theta^T$ and latest labeling of the training set $L^T = \{y_i^j = F(x_i^j, \theta^T)\}$

Algorithm 2 Evaluating expected change

Input: Training set τ, current parameters θ, current labeling $L = \{y_i^j = F(x_i^j, \theta)\}$
Output: EC scores for each latent variable

1. **for each** latent variable y_i^j
   (a) **for each** admissible label $l \in Y^j$
      i. retrain appearance model parameters $\theta' = (\theta^t | y_i^j = l)$
      ii. infer MAP labeling with unary potentials $\psi\left(y_i^j, x_i^j, \theta'\right)$ and y_i^j clamped to l
      iii. record change $C(y_i^j, l) = \sum_{i' \neq i, j' \neq j} w_{i'}^{j'} \left[F\left(x_{i'}^{j'}, \theta^t | y_i^j = l\right) \neq F\left(x_{i'}^{j'}, \theta^t\right) \right]$
   (b) set $EC(i,j) = \frac{1}{|Y^j|} \sum_{l \in Y^j} C(y_i^j, l)$
2. return EC

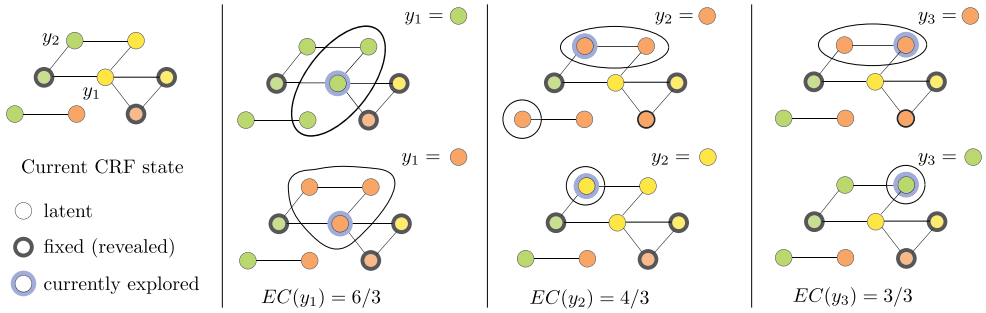


Figure 2. **Evaluating Expected Change (EC).** (Leftmost panel) Current state of the CRF. Each node is a variable and different colours represent different states (i.e. class labels). The states of nodes with thick borders are known, as previously revealed by the oracle, the others are latent. (Next three panels) EC is evaluated for three latent variables y_1, y_2, y_3 (blue halo in their respective panel). For each y, every admissible state is hypothesized in turn, and we measure the change induced by setting y to that state. The area of influence for each hypothesized state is enclosed in black (i.e. the set of variables which are changed as a consequence). Note, that disconnected nodes can influence each other via retraining of appearance models.

4.1. Fast Expected Change via dynamic graph cuts

During EC evaluation, NK runs of alpha-expansion must be performed. We explain here how to speed up this expensive step by employing dynamic graph-cuts [13].

A single run of alpha-expansion iterates through all possible labels $\mathcal{Y}$. For the current label α a binary problem is setup, where each latent variable has the choice to either retain its current label, or switch to α. This problem can be solved efficiently and exactly via graph-cuts [23], and it is guaranteed not to increase the original multilabel energy. The algorithm cyclically iterates over labels until no label can expand further.

All binary problems operate on the same graph with the same edges between nodes. This constraint holds true *beyond a single run of alpha-expansion*, over the loops over variables and labels in Alg. 2. Therefore, at each consecutive graph-cut step we *recycle* the labeling (primal solution) and the flow (dual solution) from the previously solved graph-cut, both inside an alpha-expansion run and from the previous variable/label combination in Alg. 2). As detailed in [13], this recycling involves updating the flow to render it consistent with the new problem, and then running binary graph-cut (now at a much lower cost thanks to the recycled flow). This make alpha-expansion in consecutive runs up to $12\times$ faster than in the initial run.

4.2. Batch queries

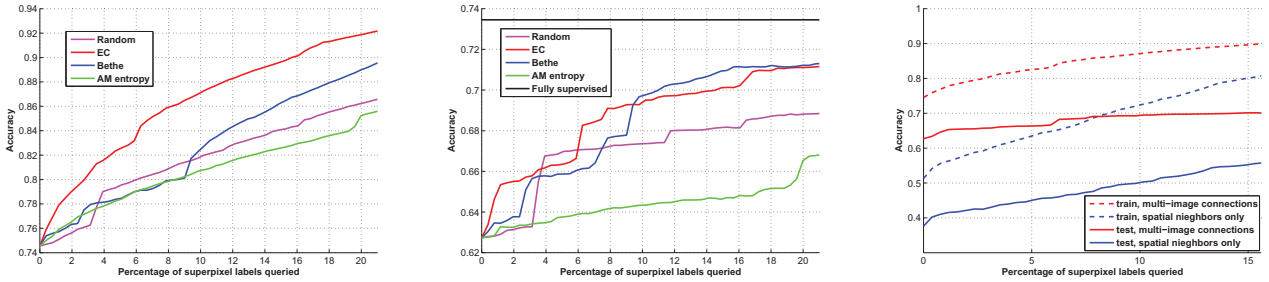
When the oracle reveals the label of a latent variable, this induces changes in the labels of other variables as well. Locally, through the pairwise potentials, and globally, through the unary potentials, due to the retraining of the appearance models. During EC evaluation (Alg. 2) we directly estimate the influence area of each latent variable y, i.e. which variables are expected to change if we query the oracle for the label of y. . The crucial observation is that often the influence areas of different variables *overlap* (fig. 2). We can exploit this fact in Alg. 1 by using a rule that queries for the labels of several variables in a single batch Q. This produces a speedup equal to the size of Q, economizing on the most expensive part of the algorithm - score computation. We aim to query labels for variables whose influence areas overlap the least, so that Q induces a maximal total expected change. Let the *expected influence area* of y_i^j be $\delta(y_i^j, \theta) = \{y_a^b : \exists l \in Y^j : F(x_a^b, \theta) \neq F(x_a^b, \theta | y_i^j = l)\}$ and $|\Upsilon| = \sum_{a,b:y_a^b \in \Upsilon} w_a^b$. We want to find the set of queries Q that maximizes expectation of expected influence area of the query: $|\bigcup_{y_i^j \in Q} \delta(y_i^j, \theta)|$. Direct maximization is NP-hard as this is an instance of the set cover problem [27]. Therefore we use a greedy approximation. We start with Q containing only the variable with the highest EC score. We then add the next highest-scored variable which does not belong to the influence area of any variable in Q. The process is repeated until a predefined number of queries is selected.

In our experiments, using dynamic graph cuts and batch queries bring a combined speedup of about factor 1200x.

5. Experimental results

We evaluate the proposed methods on two well known data-sets for semantic segmentation: MSRC-21 [5] and the subset of LabelMe defined by [14].

Image features. We describe the appearance of superpixels with a bag of semantic textons (BoST) [4] trained from weakly supervised data using [7]. BoST enables to represent superpixels as histograms, on which we train a linear regressor to predict class labels (appearance model θ, sec. 3). It also enables to use as a similarity $D(x_i^j, x_{i'}^{y'})$ Histogram Intersection [4] in the pairwise potential (eq. (3)). We scale ϕ by median of maximum per superpixel contribution of all pairwise potentials to energy in MIM to make



(a) Weakly labeled training set (b) Test set (c) With and without multi-image connections

Figure 3. Results on MSRC data-set. Plotted is the accuracy over share of queries asked: a) accuracy on weakly supervised training set; b) accuracy on test set; c) comparison between a CRF with only pairwise connections between spatial neighbours within one image vs also including multi-image connections.

them comparable to unary potentials.

Baselines. We compare to three competing criteria for scoring queries: i) a baseline random sampling, ii) the uncertainty sampling by the entropy of the unary potential (this score is calculated from the outputs of Naive Bayes appearance models). iii) the third criterion samples according to maximal uncertainty measured by the Bethe entropy [28] of the full CRF. It was used in [10] for active learning for semantic segmentation, which makes it a very relevant baseline. Bethe entropy is a more sophisticated technique, which takes the CRF connectivity into account and it approximates the full entropy of the CRF (more details in [28, 10]). Computing Bethe involves looping over i) all pairs of nodes, and ii) all possible labellings of the pair to compute partition sums. This makes this criterion expensive to compute, thus recomputing the score for each node after every query is prohibitive. To make it feasible we query top N nodes instead of one, after computing the score for each unknown node. We match N, such that wall time of Bethe and EC are, approximately, the same. With batch queries disables for both EC and Bethe, we observed EC being approximately only 25% slower. Notice that Bethe does not allow for a principled batch query as our method does (Sec. 4.2).

All criteria are embedded in the same Alg. 2. For EC and Bethe the sizes of batch queries are 0.4% and 0.3% of the variables respectively. For all data-sets, we query until 97% of the performance of the fully supervised model on the test set is reached by the best method.

Generalizing to test data. After the learning stage has recovered the labels of the superpixels in the training images (sec. 4), we can train any standard fully supervised method and then employ it to label a new test image T. In our experiments, we use a method from [6]. First, we retrieve the most similar training images to T, using a pre-trained multiple kernel metric [29]. Using this metric, we predict image-level label probabilities for T, called *image-level prior (ILP)* in [4]. Then the following energy is minimized to find the optimal labeling of the superpixels y_i^t of T:

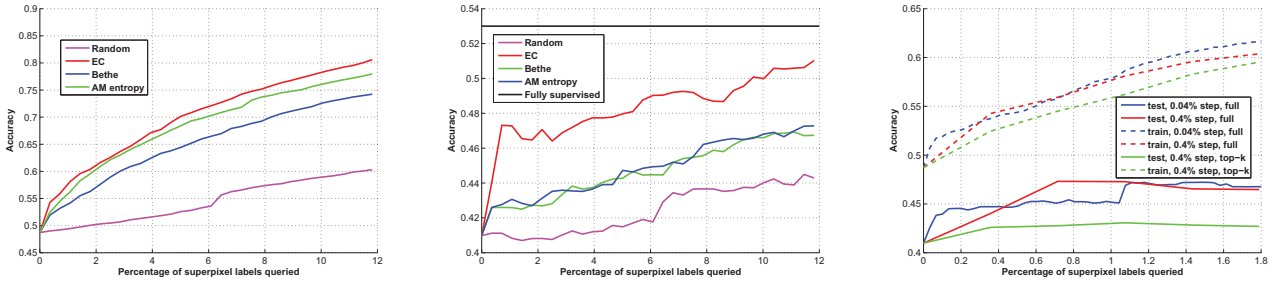
$$\mathcal{E}\left(\{y_i^t\}\right) = \sum_i \left(\psi\left(y_i^t, x_i^t, \theta^*\right) + \mu\left(y_i^t, I^t\right)\right) + \sum_{(y_i^t, y_{i'}^j) \in S} \phi\left(y_i^t, y_{i'}^t, x_i^t, x_{i'}^t\right) + \sum_{(y_i^t, y_{i'}^j) \in M^t} \phi\left(y_i^t, y_{i'}^j, x_i^t, x_{i'}^j\right) \tag{6}$$

The first unary potential $\psi\left(y_i^t, x_i^t, \theta^*\right)$ measures how well x_i^t matches the appearance model of class $\theta^*_{y_i^t}$ (note the parameters θ^* are fixed as learnt during training). The second unary potential is ILP, which can be seen as a soft version of π in eq. (1). Pairwise potentials ϕ connect neighbouring superpixels in the test image (set S). We also connect each test image superpixel y_i^t to its 3 most similar superpixels in each retrieved training image (set M). Note how variables y_i^j from the training images are fixed, which facilitates optimization.

MSRC-21. This popular data-set [5] contains 591 images of 320x213 pixels, accompanied by ground-truth segmentations of 21 classes [1]. We use the standard split into 276 training and 256 test images defined by [5]. This data-set is best suited for our task, as all classes are labeled in all images and there exist significant co-occurrence between classes.

Fig. 3a reports the accuracy in recovering superpixel labels on the weakly labeled training set, as a function of the percentage of queries asked to the oracle. Our newly proposed EC criterion performs considerably better than all competing criteria we compare to, over the whole range of percentage of queries. Until 10% queries are asked, the competing criteria perform approximately equally, with Bethe entropy sampling taking the lead afterwards. As fig. 3b shows, on the test set EC and Bethe are clear leaders, and they converge to the same result after 17% queries. Surprisingly, random sampling performs better than the entropy

[1] Two additional classes in this data-set (mountain and horse) are usually discarded as they occur very rarely.



(a) Weakly labeled training set (b) Test set (c) Batch query variations

Figure 4. Results on LabelMe data-set. Plotted is the accuracy over share of queries asked: a) accuracy on weakly supervised training set; b) accuracy on test set; c) variations of the batch query scheme: our full scheme from sec. 4.2 (0.4%, full), a scheme with a 10× smaller batch size (0.04%), and a simplified scheme where the top-k superpixel labels are queried, without taking into account their influence areas.

of the unary potential. This effect might be due to the latter method neglecting the CRF connections, thereby overfitting to unary potential uncertainty. Moreover, the good performance of the random strategy suggests this data-set to be fairly simple. As a reference, we also plot the test accuracy of a fully supervised method (equivalent to asking 100% queries). Both EC and Bethe reach 97% of its accuracy after querying 17% of the training data. Note the sudden jumps in test accuracy. These discontinuous improvements correspond to critical amounts of information where the learning algorithm has "understood" special subclasses, like grey sky or yellowish grass.

LabelMe [14]. The LabelMe subset of [14] contains 2500 images with 34 classes and it is more challenging than MSRC-21. For computational reasons, we train on subset of 800 images of the training set defined by [14] (but we use the same test set as [14]). As in MSRC-21, all classes are labeled in all images and there is significant co-occurrence between classes. All learning procedure parameters are kept the same as for MSRC-21.

As fig. 4 shows, EC performs better than all other methods on the weakly labeled training set, as on MSRC-21. However, different than on MSRC-21, EC is clearly the best method also on the test set. Sampling schemes based on unary potential entropy and Bethe entropy perform about the same on the test set. Random sampling performs much worse then other criteria, confirming our judgement that this data-set is more challenging.

Evaluation of components. Here, we present an experiment to evaluate the contribution of pairwi potentials connecting superpixels between different images in the CRF model for the training set. Generalization to the test set is kept same (sec. 5). Fig. 3(c) plots results of our EC criterion on MSRC-21, using connections E including only spatial neighbours within an image, vs also including multi-image connections (sec. 3). The initial accuracy (pure weakly supervised learning) is significantly lower without multi-image connections. The learning rate is faster though, which is natural as both of them will eventually converge to the accuracy of a fully supervised method. The difference emphasizes the benefits of exploiting the hidden dependencies between distant superpixel labels.

In fig. 4(c) we evaluate variations on our batch query scheme (sec. 4.2). First, we use a 10× smaller batch size (i.e. 0.04curves (red and blue) are very close, conrming that the proposed batch query scheme considerably accelerates active learning without compromising accuracy. Another variation is a simplied scheme which takes the top-k queries based only of their EC score without taking inuence areas into account. Its performance is signicantly worse, which demonstrate the value of selecting queries affecting different regions of the CRF, as in our full scheme.

6. Conclusion

We presented an exploration of the gap between weakly and fully supervised methods for semantic segmentation by active learning. For this purpose, we introduced a novel *Expected Change* score of the informativeness of nodes in a pairwise CRF model. High computational complexity is being remedied by application of dynamic graph cuts and principled batch query strategy. Our method consistently outperforms relevant baselines, including Bethe entropy sampling, on both an easy and a difficult data-set, having the same wall time speed as Bethe. It reaches 97% of total pixel accuracy of the corresponding fully supervised model, while querying less than 17% of the superpixel labels. The experiments reveal the existence of strong and far-reaching hidden dependencies in semantic segmentation data. Exploiting those dependencies enables to significantly reduce the supervision effort. Finally, our method could be used for other problems that can be formulated as a pairwise CRF estimation.

Acknowledgments

A. Vezhnevets was supported by the SNSF under grant #200021-117946. V. Ferrari was supported by a SNSF Professorship.

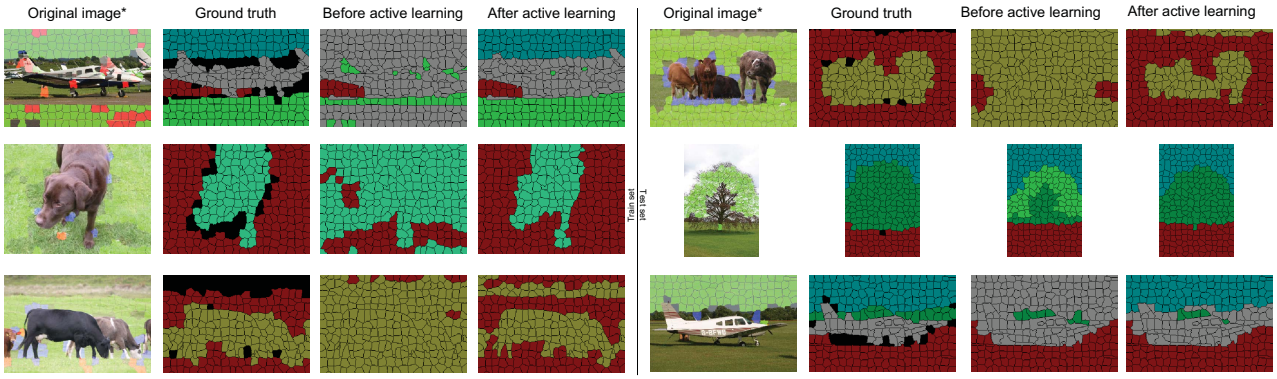


Figure 5. **Semantic segmentation results on MSRC data-set.** Columns 1-4 present results on training set, column 5-8 on test set. From left to right, images present (i) original image (ii) ground truth, (iii) semantic segmentation before active learning (only weakly supervised learning) (iv) after 6.8% of labels queried using EC criterion. In original images superpixels highlighted by green changed their labels to the correct one after active learning, highlighted by blue changed their label to the wrong one; on training set, red highlights those superpixels that had their label queried. Notice, that in training images a lot of change happens with only few labels being queried. This is achieved by directly targeting those queries, that produce maximum expected change. In row two, training image has most of its superpixel labels changed, although only one of its superpixel labels has been queried. This is due to propagation of change through unary and multi-image potentials.

References

[1] L. Ladicky, C. Russell, P. Kohli, and P. Torr, "Graph cut based inference with co-occurrence statistics.," in *ECCV*, 2010. 1, 2, 3

[2] L. Ladicky, C. Russell, and P. Kohli, "Associative hierarchical crfs for object class image segmentation," in *CVPR*, 2009. 1, 2, 3

[3] P. Kohli, L. Ladicky, and P. Torr, "Robust higher order potentials for enforcing label consistency.," in *CVPR*, 2008. 1, 2, 3

[4] J. Shotton, M. Johnson, and R. Cipolla., "Semantic texton forests for image categorization and segmentation," in *ECCV*, 2008. 1, 2, 5, 6

[5] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *ECCV*, 2006. 1, 2, 3, 5, 6

[6] A. Vezhnevets, V. Ferrari, and J. Buhmann, "Weakly supervised semantic segmentation with a multi image model," in *ICCV*, 2011. 1, 2, 3, 6

[7] A. Vezhnevets and J. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning.," in *CVPR*, 2010. 1, 2, 5

[8] J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in *CVPR*, 2007. 1, 2, 3

[9] S. Vijayanarasimhan and K. Grauman, "What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations," in *CVPR*, pp. 2262–2269, 2009. 1, 2, 4

[10] B. Siddiquie and A. Gupta, "Beyond active noun tagging: Modeling contextual interactions for multi-class active learning.," in *CVPR*, 2010. 1, 2, 4, 6

[11] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active learning with gaussian processes for object categorization," in *ICCV*, pp. 1–8, 2007. 1, 2, 4

[12] A. J., F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *CVPR*, 2009. 1, 2, 4

[13] P. Kohli and P. H. S. Torr, "Dynamic graph cuts for efficient inference in markov random fields," *TPAMI*, vol. 29, no. 12, pp. 2079–2088, 2007. 1, 5

[14] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: label transfer via dense scene alignment.," in *CVPR*, 2009. 2, 5, 6

[15] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," in *CVPR*, 2011. 2

[16] S. Vijayanarasimhan and K. Grauman, "Multi-level active prediction of useful image annotations for recognition," in *NIPS*, 2008. 2, 4

[17] A. Krause and C. Guestrin, "Optimal value of information in graphical models," *J. Artif. Intell. Res. (JAIR)*, pp. 557–591, 2009. 2

[18] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *UAI*, July 2005. 2

[19] B. Anderson and A. W. Moore, "Fast information value for graphical models," in *NIPS*, 2005. 2

[20] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *NIPS*, pp. 1289–1296, MIT Press, 2007. 2

[21] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *TPAMI*, vol. 31, pp. 2290–2297, 2009. 2

[22] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," in *ECCV*, 2002. 3

[23] Y. Boykov and M.-P. Jolly, "Cuts for optimal boundary and region segmentation of objects in n-d images.," in *CVPR*, 2001. 3, 5

[24] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts.," *TPAMI*, 2001. 3

[25] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision.," *TPAMI*, 2004. 3

[26] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Machine Learning*, pp. 209–239, 2004. 3

[27] R. Karp, "Reducibility among combinatorial problems," in *Complexity of Computer Computations* (R. Miller and J. Thatcher, eds.), pp. 85–103, Plenum Press, 1972. 5

[28] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms.," *IEEE Transactions on Information Theory*, pp. 2282–2312, 2005. 6

[29] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *ICCV*, pp. 309–316, sep 2009. 6