Learning to Segment from a Few Well-Selected Training Images

Alireza Farhangfar FARHANG@CS.UALBERTA.CA
Russell Greiner GREINER@CS.UALBERTA.CA
Csaba Szepesvári SZEPESVA@CS.UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, T6G 2E8

Abstract

We address the task of actively learning a segmentation system: given a large number of unsegmented images, and access to an oracle that can segment a given image, decide which images to provide, to quickly produce a segmenter (here, a discriminative random field) that is accurate over this distribution of images. We extend the standard models for active learner to define a system for this task that first selects the image whose expected label will reduce the uncertainty of the other unlabeled images the most, and then after greedily selects, from the pool of unsegmented images, the most informative image. The results of our experiments, over two real-world datasets (segmenting brain tumors within magnetic resonance images; and segmenting the sky in real images) show that training on very few informative images (here, as few as 2) can produce a segmenter that is as good as training on the entire dataset.

1. Introduction

Many imaging tasks involve segmentation. For example, given a Magnetic Resonance (MR) image of the brain, it is important to find and segment any tumor region present. Many effective imaging systems involve a number of parameters that have to be adjusted; some of these systems therefore include a *learning* component that can learn effective parameters from a set of labeled (that is, segmented) images. In general, these systems require a large number of such labeled images to produce an effective segmenter. Fortunately, there are often a large number of available images — perhaps on the web, or in clinical databases. Unfortunately, most such images are unlabeled, and worse, it can be expensive to obtain the labels (as this may require paying a medical doctor to label each image, which is costly in terms of both time and money). This often limits the amount of training data available, which can lead to an inferior segmentation system.

An *active learning process* tries to address this problem by identifying which of the unlabeled images should be labeled. Section 2 overviews this body of work, to help motivate our approach. Section 3 then presents our actual active learning algorithm, LMU. It also describes how the underlying performance system — here using a Discriminative Random Field — segments the images. Section 4 shows the results of our experiments using this active learner on two real-world datasets: Finding the sky in the Geometric Context dataset (Szummer et al., 2008) and segmenting tumors within MR images of human brains (Lee et al., 2008). In particular, we compare the segmentation performance (on hold-out images) of a segmenter trained on all labeled images, versus one trained using only the first k images selected by our active learner, for various values of k. (We also consider segmenters learned from k randomly-selected images.) We find that, surprisingly, the segmenter based on only $k = 2$ well-selected images is typically as good as the one based on *all* (here, 85) images! But only if the images are well-selected; the segmenter based on $k = 2$ *random* images is typically considerably inferior, as is one based on a larger number of random images. (See also our webpage (Greiner et al., 2009) for additional information — *e.g.,* timing information, studies using the simpler CalTech101 images (Fei-Fei et al., 2007), etc.)

2. Related Work

Most supervised learning systems are passive, in that they produce a classifier based only on the existing corpus of labeled training instances. By contrast, an *active learner* is able to extend this set of labeled instances by sequentially identifying an unlabeled instance and obtaining its label from an oracle, then adding the resulting *labeled* instance to the training data.

There are many results on active learning, most of which

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

relate to the standard supervised learning framework, where the label for an instance $\mathbf{x} \in \mathbf{X}$ is drawn from a small set of possible labels $y \in Y$ (*e.g.,* Y ={+1, -1}). Some of these algorithms select the instance $\mathbf{x}$ whose label y is most uncertain, based on the $\hat{P}^{(D)}(y|\mathbf{x})$ probability distribution obtained by the current training data D. Freund *et al.* (Freund et al., 1997) select the instance that has maximal disagreement from the current committee of classifiers. Many researchers, working with support vector machines, select the instance closest to the boundary (Tong & Koller, 2002; Schohn & Cohn, 2000; Campbell et al., 2000). Lewis and Gale (Lewis & Gale, 1994) use a probabilistic classifier, and select the most uncertain instance — *i.e.,* the one whose conditional probability of +class given the features is closest to 0.5, assuming binary labels. Our LMU system (Section 3.2) uses a segmentation-analogue to this basic approach as part of its process.

This most uncertain approach typically works well if the current parameters nicely approximate the conditional probability distribution of the entire data. Unfortunately, these parameters are often problematic as they are based on a very small training set. As the goal is to learn a classifier that works well on the distribution over $\mathbf{X}$, it makes sense to consider the marginal distribution over unlabeled data $P(\mathbf{X})$. This motivates a second class of approaches that use the pool of unlabeled instances. Some active learners use clustering algorithms to first group the unlabeled data. Nguyen and Smeulders (Nguyen & Smeulders, 2004) repeatedly cluster the unlabeled data and then request labels of one representative from each cluster. Xu *et al.* (Xu et al., 2003) request labels for the instances near the centers of cluster lying within the margin of support vector machine. Other systems explicitly use P(X); eg, Cohn *et al.* (Cohn et al., 1996) and Zhang and Chen (Zhang & Chen, 2002) estimated and then used the density $P(\mathbf{X})$ as weights for unlabeled data. Roy and McCallum (Roy & McCallum, 2001) selected instances that reduce expected error over unlabeled data. Guo and Greiner (Guo & Greiner, 2007) proposed an algorithm that selects the instance that provides the maximum conditional mutual information about the labels of unlabeled instances. Our LMU system also incorporates (a segmentation-analogue to) this idea, for its first iteration.

The underlying domain of $\mathbf{X}$ is quite arbitrary, ranging from simple binary tuples to feature sets obtained from natural language texts. Some active learners deal with images, which are complicated due to their large dimensionality. Vijayanarasimhan and Grauman (Vijayanarasimhan & Grauman, 2008) proposed a framework to actively recognize objects inside the image (from a small predefined set of labels), using a mixture of weakly and strongly labeled images. Their method selects the partially labeled or unlabeled instance that minimizes the expected risk of other instances. The system by Collins *et al.* (Collins et al., 2008) actively selects the most uncertain instance, then uses boosting techniques to train a decision stump classifier to detect and recognize various objects.

As noted above, most of the existing systems have been used primarily to learn a classifier that maps each instance to a *simple label*; even the imaging work mentioned above has focused on mapping each image to one of small set of labels Y. As images can often be recognized based only on a small set of extracted features, object recognition corresponds to a typical machine learning problem. There have been relatively little active learning research related to *image segmentation*, which requires producing a more complicated label: such systems map an image of $n \times m$ pixels to $n \times m$ individual (correlated) pixel-labels; if each pixel-label is binary, this means that $Y = \{+1, -1\}^{n \times m}$. Moreover, these pixel-labels are not independent of one another (*e.g.,* if one pixel is a tumor, it is more likely that its neighbors are, as well). This forces a segmenter to consider the entire $n \times m$ image as an instance, rather than label one pixel at a time. Hence, notions like uncertainty and information content must be defined for the entire image.

3. Implementation

This section overviews our basic system. We first describe the training and performance of the underlying segmentation system, based on Discriminative Random Fields (DRFs) (Kumar & Hebert, 2003). (A DRF is a version of a conditional random field (CRF) that is designed to deal with variables that are organized in a 2-dimension grid.) We then summarize our LMU system, which actively learns the parameters for the DRF.

For notation, we let $\mathbf{I}$ represent the set of $n \times m$ image pixels, $\mathbf{x} = \{x_i \mid i \in \mathbf{I}\}$ be an observed input image, where each x_i is a vector describing pixel i (perhaps its intensity and texture) and $\mathbf{y} = \{y_i \mid i \in \mathbf{I}\}$ is the corresponding joint set of labels over all pixels of the image. We will assume the segmentation is *binary* over $n \times m$ images, where each $y_i \in \{-1, +1\}$ (*e.g.,* is this pixel tumorous vs healthy), and the overall output $\mathbf{y}$ is $n \times m$ such bits.

At any time, our system has a pool of unsegmented instances U, as well as a (possibly empty) pool of segmented instances L.[1] Our active learner will sequentially select an unsegmented image $u \in U$, obtain its label (recall this label is a set of $|u|$ bits — one for each pixel in u), and then move this now-labeled image from U to L.

[1] Later, when we describe our experiments, we will also have a set of unlabeled test images T. Note that these sets, T, L, U, are disjoint. Here, and below, we will use labeled as a synonym for segmented and unlabeled for unsegmented.

3.1. Discriminative Random Field

A discriminative random field (DRF) is a model of the conditional probability of a set of labels $\mathbf{y}$ given the observations $\mathbf{x}$, here given by

$$P_\theta(\,\mathbf{y}\,|\,\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})}\exp\left(\sum_{i\in I}\Phi_{\mathbf{w}}(y_i,\mathbf{x}) + \sum_{i\in I}\sum_{j\in N_i}\Psi_{\mathbf{v}}(y_i,y_j,\mathbf{x})\right) \quad (1)$$

where N_i are the 4 neighbors of i (north, east, south, west); $\Phi_w(y_i,\mathbf{x}) = \log(\frac{1}{1+\exp(-y_i\mathbf{w}^T h_i(\mathbf{x}))})$ is the *association potential* of pixel i that uses the association parameters $\mathbf{w}$ obtained from training data, and $h_i(\mathbf{x})$ is the feature vector for pixel i;[2] $\Psi_v(y_i,y_j,\mathbf{x}) = y_i y_j \mathbf{v}^T\mu_{ij}(\mathbf{x})$ is the interaction potential that captures the spatial correlation with neighboring pixels using the interaction parameters $\mathbf{v}$ obtained from the training data; $\mu_{ij}(x) = x_i - x_j$ is the difference between feature vectors in pixel i and j; and $\theta = [\mathbf{w},\mathbf{v}]$ are the model parameters. The normalizing factor

$$Z_\theta(\mathbf{x}) = \sum_y \exp\left(\sum_{i\in I}\Phi_{\mathbf{w}}(y_i,\mathbf{x}) + \sum_{i\in I}\sum_{j\in N_i}\Psi_{\mathbf{v}}(y_i,y_j,\mathbf{x})\right) \quad (2)$$

insures that the DRF produces a probability.

Given a set of parameters θ, we can compute the label $\mathbf{y}^*$ for a given image $\mathbf{x}$: $\mathbf{y}^* = \arg\max_{\mathbf{y}} P_\theta(\,\mathbf{y}\,|\,\mathbf{x})$ (see Section 3.1.2). The challenge is learning the best values for these parameters, θ^*. The rest of this subsection discusses how to compute these optimal parameters from a set of labeled images L and then for this L and a single unlabeled image u; it then provides a useful approximation to this computation, to avoid its inherent intractability.

3.1.1 Training: Typical supervised DRF training involves finding the parameters

$$\theta^*_L = \arg\max_\theta \mathrm{RL}_L(\,\theta\,) \quad (3)$$

that maximize the log of the posterior probability over training set of labeled images L

$$\mathrm{RL}_L(\,\theta^*\,) \;=\; \sum_{\ell\in L}\log P_{\theta^*}(\,\mathbf{y}^{(\ell)}\,|\,\mathbf{x}^{(\ell)})$$

In our active learning framework, we need to consider the effect of adding one more image from unlabeled set to the training set. Since the label $y^{(u)}$ of an unlabeled image $x^{(u)}$ is not known before presenting it to an expert, we use a conditional entropy term to express the likelihood associated with the unlabeled image as

$$\mathrm{RL}_u(\,\theta\,) \;=\; \sum_{\mathbf{y}} P_\theta(\,\mathbf{y}\,|\,\mathbf{x}^{(u)})\log P_\theta(\,\mathbf{y}\,|\,\mathbf{x}^{(u)})$$

The overall objective function now consists of two terms: the first term deals with all labeled images in training set L and the second term is for an unlabeled image u.

$$\mathrm{RL}_{L+u}(\,\theta\,) \;=\; \mathrm{RL}_L(\,\theta\,) \;+\; \gamma\,\mathrm{RL}_u(\,\theta\,) \quad (4)$$

where the $\gamma \in \Re$ parameter trades-off the two factors. We then seek the parameters θ^*_{L+u} that maximize Equation 4.

3.1.2 Useful Approximation: Given the lattice neighborhood structure of the DRF, it is intractable to compute the normalizing factor $Z_\theta(\mathbf{x})$ in Equation 2. Following (Kumar & Hebert, 2003), we therefore incorporate the pseudo-likelihood approximation, which assumes that the joint probability distribution of all pixel labels $\mathbf{y}$ can be approximated by the product of "local probabilities" of each pixel, which is based on only the observations of x_i and the labels of the neighboring nodes y_{N_i}:

$$\begin{aligned}\hat{P}_\theta(\mathbf{y}|\mathbf{x}) &\approx \prod_{i\in I}\hat{P}_\theta(y_i|y_{N_i},\mathbf{x})\\ \hat{P}_\theta(y_i|y_{N_i},\mathbf{x}) &= \frac{1}{z_i(\mathbf{x})}\exp\,(\,\Phi_{\mathbf{w}}(y_i,\mathbf{x}) + \sum_{j\in N_i}\Psi_{\mathbf{v}}(y_i,y_j,\mathbf{x}))\end{aligned} \quad (5)$$

where this $z_i(\mathbf{x})$ is a "local normalizing term", which deals only with the i^{th} pixel.

Using the approximation in Equation 5, the entropy regularization term for (respectively) training set L and a single unlabeled image u is:

$$\widehat{\mathrm{RL}}_L(\,\theta\,) \;=\; \sum_{\ell\in L}\sum_{i\in\mathbf{I}^{(\ell)}}\log\hat{P}_\theta(y_i^{(\ell)}|y_{N_i}^{(\ell)},x_i^{(\ell)}) \quad (6)$$

$$\widehat{\mathrm{RL}}_u(\,\theta\,) \;=\; \sum_{i\in\mathbf{I}}\sum_{y_i}\hat{P}_\theta(y_i|y_{N_i}^{(u)},x_i^{(u)}) \;\times\; \log\hat{P}_\theta(y_i|y_{N_i}^{(u)},x_i^{(u)}) \quad (7)$$

Following Equation 4, we can combine these to form a single objective that combines the effect of the many labeled images L and the one unlabeled data u:

$$\widehat{\mathrm{RL}}_{L+u}(\,\theta\,) \;=\; \widehat{\mathrm{RL}}_L(\,\theta\,) \;+\; \gamma\,\widehat{\mathrm{RL}}_u(\,\theta\,) \quad (8)$$

[2] Here, this $h_i(\mathbf{x})$ is just the information in x_i. In general, it could also include information from some adjacent pixels, via some smoothing operator, etc.

In our experiments, we set $\gamma = 1$. We also used conjugate gradient to optimize the objective function in Equation 8.

The $\widehat{\mathrm{RL}}_u(\,\theta\,)$ term from Equation 7 requires the label for unlabeled data that is not yet available, *i.e.,* $y_{N_i}^{(u)}$ is not known. An inference step is added to estimate the labels based on current parameters. The inference is based on iterative conditional probability (ICM) (Besag, 1986), which sets the label for pixel as the maximum posterior probability:

$$y_i^* = \arg\max_{y_i} P_\theta(\,y_i \,|\, y_{N_i}, \mathbf{x})$$

where for each pixel i, we assume that the labels of its neighbors y_{N_i} are fixed to their current estimate. We then use y_i^* to compute the label for its neighbors. We repeat this process until every y_i converges to its final value.

3.2. The LMU Active Learning Algorithm

At each time, given U and L, our LMU active learner needs to select which unlabeled image $u \in U$ to give to the oracle for labeling. (Recall that this now-labeled u will then be added to the set of labeled images L.) One option is to choose the most uncertain image:

$$\mathrm{MU}(U, L) \quad = \quad \arg\max_{u \in U} H(\,\mathbf{Y}^{(u)} \,|\, \mathbf{x}^{(u)}, L\,)$$

where

$$H(\,\mathbf{Y} \,|\, \mathbf{x}, L\,) = -\sum_{\mathbf{y} \in \mathbf{Y}} P(\,\mathbf{y} \,|\, \mathbf{x}, L\,) \log P(\,\mathbf{y} \,|\, \mathbf{x}, L\,) \quad (9)$$

$$\approx -\sum_{\mathbf{y} \in \mathbf{Y}} \hat{P}_{\theta_L}(\mathbf{y}|\mathbf{x}) \log \hat{P}_{\theta_L}(\mathbf{y}|\mathbf{x}) \qquad (10)$$

approximates the conditional entropy of the label $\mathbf{Y}$ given the image observations $\mathbf{x}$, based on the current conditional probability, which is based on the training data L. (To explain Equation 10: As we use the data L to produce the parameters θ_L, we identify $P(\,\mathbf{y} \,|\, \mathbf{x}, L\,)$ with $\hat{P}_{\theta_L}(\mathbf{y}|\mathbf{x})$; see Equation 3 and relevant approximation.)

The summation in Equation 10 is over all $|Y| = 2^{m \times n}$ possible binary assignments to $n \times m$ pixels. We approximate this as the simple sum of the entropies of the labels y_i of each pixel of the image, $i \in \mathbf{I}$:

$$\hat{H}(\,\mathbf{Y} \,|\, \mathbf{x}, L\,) \;=\; \hat{H}(\,\mathbf{Y} \,|\, \mathbf{x}, \theta_L\,) \;=\; -\sum_{i \in \mathbf{I}} \sum_{y_i \in \{\pm 1\}} \left[\hat{P}_{\theta_L}(y_i|x_i) \log \hat{P}_{\theta_L}(y_i|x_i)\right] \qquad (11)$$

We view this most-uncertain instance $\mathrm{MU}(U, L)$ as being at the *boundary* between positives and negative instances. Having the label for such boundary points will help us to define the boundary more precisely and consequently increase the classification accuracy. Of course, our active learner has to start with an empty $L = \{\}$; here the associated $\theta_{\{\}}$ parameters are problematic. Moreover, this approach does not consider the distribution over $\mathbf{X}$, which means knowing more about this "boundary" point might not help identify that much wrt this $\mathbf{X}$.

This suggests an alternative approach: select the instance that would provide the maximum information about the labels of remaining unlabeled instances — *i.e.,* the instance that most reduces the uncertainty (RU) of the other unlabeled images $U - \{u\}$:

$$\mathrm{RU}(U, L) = \arg\max_{u \in U} H(\,\mathbf{Y}_U \,|\, \mathbf{X}_U, L\,) - H(\,\mathbf{Y}_U \,|\, \mathbf{X}_U, L + \mathbf{x}^{(u)}\,) \qquad (12)$$

$$= \arg\min_{u \in U} H(\,\mathbf{Y}_U \,|\, \mathbf{X}_U,\; L + \mathbf{x}^{(u)}\,) \qquad (13)$$

$$= \arg\min_{u \in U} \sum_{v \in U; v \neq u} H(\,\mathbf{Y}^{(v)} \,|\, \mathbf{x}^{(v)},\; L + \mathbf{x}^{(u)}\,) \qquad (14)$$

$$\approx \arg\min_{u \in U} \sum_{v \in U; v \neq u} \hat{H}(\,\mathbf{Y}^{(v)} \,|\, \mathbf{x}^{(v)},\; \theta_{L + \mathbf{x}^{(u)}}\,) \qquad (15)$$

Equation 13 follows from the observation that the first term of Equation 12 is constant for all instances u; Equation 14 uses the fact that the entropy of the set of independent images is just their sum; and Equation 15 again uses the approximate conditional entropy $\hat{H}$ defined in Equation 11. (This $\theta_{L + \mathbf{x}^{(u)}}$ is the solution to Equation 8.)

Note this RU approach works even for $L = \{\}$; of course, it is computationally more expensive than MU approach.

Our actual LMU system uses both approaches: Given a set of unlabeled images U and no labeled instances $L = \{\}$, it first uses RU to find the first instance u_1 to label, then sets $L = \{u_1\}$. Thereafter, it used MU to find the second, third, and further images. See Figure 1.

4. Experiments

To investigate the empirical performance of our active learning algorithm, we conducted a set of experiments on two challenging real-world problems: Finding the sky in the geometric context dataset and segmenting tumors in medical images. We also ran a scaling study, to see the influence of the size of each image on the number of images required to obtain good performance.

4.1. Finding the Sky in Color Images

The geometric context dataset (Szummer et al., 2008) is a collection of 125 images, many very cluttered, that span a variety of natural, urban, and suburban sceneries. Our goal here is to find the sky within these images. Figure 2 shows three instances of images in this dataset. This task is challenging since the sky could be blue or white, clear or overcast, and worse, many scenes contain both sky and ocean,

```
LMU( $U$: unsegmented images )
$L := \{\}$        % $L$ is initially empty
% Compute $u_1 = RU(U, \{\})$
for each unlabeled image $u \in U$
    $\theta_u := \arg\max_\theta \widehat{\mathrm{RL}}_u(\theta)$        % Equation 7
    $s_u := 0$
    for each other unlabeled image $v \in U, v \neq u$
        $s_u \mathrel{+}= \hat{H}(\mathbf{Y}^{(v)} \mid \mathbf{x}^{(v)}, \theta_u)$        % Equation 11
$u_1 := \arg\min_u s_u$
$\mathbf{y}^{(u_1)} := \mathrm{Oracle}(\mathbf{x}^{(u_1)})$        % Get label
$L := (\langle \mathbf{x}^{(u_1)}, \mathbf{y}^{(u_1)} \rangle)$
$\theta_L := \arg\max_\theta \widehat{\mathrm{RL}}_L(\theta)$        % Equation 6
$U := U - \{u_1\}$
for $i = 2, \ldots$
    % $u_i := MU(U, L)$
    for each $u \in U$
        $t_u = \hat{H}(\mathbf{Y}^{(u)} \mid \mathbf{x}^{(u)}, \theta_L)$        % Equation 11
    $u_i := \arg\max_u t_u$
    $\mathbf{y}^{(u_i)} := \mathrm{Oracle}(\mathbf{x}^{(u_i)})$        % Get label
    $L := L + (\langle \mathbf{x}^{(u_i)}, \mathbf{y}^{(u_i)} \rangle)$
    $\theta_L := \arg\max_\theta \widehat{\mathrm{RL}}_L(\theta)$        % Equation 6
    $U := U - \{u_i\}$
end
```

Figure 1. Pseudo code for the LMU active learning algorithm

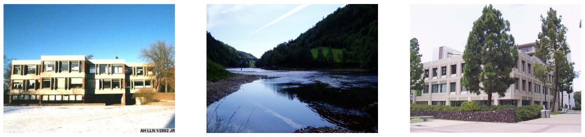

Figure 2. Sample "sky" images from Geometric context dataset

which are not easily separable based on their color. The original images were of various sizes; we downsized each to 65 × 65 pixels to make them uniform, and to make our computations more tractable. We partitioned these images into the unlabeled-set U with 85 images and the test-set T with 20 images.

Features: We associate each pixel with twelve values: its 3 color intensities, its vertical position (as the sky is typically at the top of the image) and 8 texture values: We apply the MR8 filter banks (Varma & Zisserman, 2002) (which contain filters at 6 different orientations, at 3 scales) to the region centered on each pixel, but record only the maximum filter response at each of the 6 orientations; we also include 2 isotropic features: Gaussian and a Laplacian of Gaussian.

In each iteration of the active learning process, our LMU system identifies one specific image u from U, which is removed from U, labeled by an oracle, then added to the labeled training set L. We then train a segmenter on this augmented training set $L+$"labeled"u to produce θ_{L+u}, then test this system on 20 images in the test-set T, recording the average F-measure

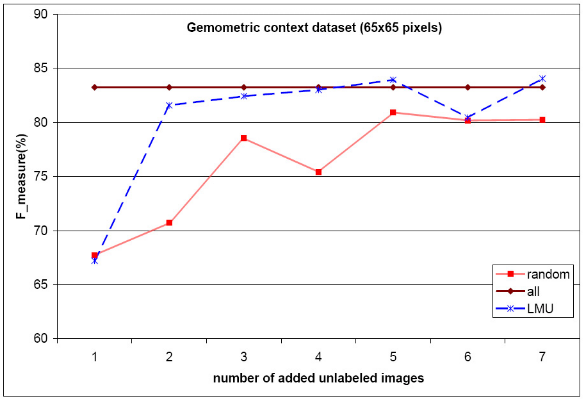

Figure 3. Accuracy of segmentation versus number of training data chosen from unlabeled set for geometric context dataset

$$\text{F-measure} \quad = \quad \frac{2 \times (precision)(recall)}{(precision + recall)}$$

where $precision = tp/(tp + fp)$ and $recall = tp/(tp + fn)$, where tp, fp, fn are true positive, false positive and false negative. Note this measure is problematic when $tp = 0$; see Footnote 3.

In Figure 3, the horizontal axis represents the number of added unlabeled images and vertical axis is the average F-measure. The "all" line shows the results of training on (the oracle-labeled versions of) *all* 85 unlabeled images. Checking the "LMU" line, we see that the first carefully-selected image alone produced a classifier whose F-measure was 67%, and this accuracy improved to 81% by using the second image. Note this is only 2% below the accuracy obtained by training on all unlabeled data; moreover, this is statistically indistinguishable at the $p < 0.05$ level, based on a paired t-test. (The accuracy of the second iteration was significantly better than the first — paired t-test $p < 0.05$.) There is no significant change between the second iteration and subsequent iterations, which means that the first two images chosen by LMU are good enough to train the classifier. Those two images, by the way, are the left and middle images shown in Figure 2.

Of course, it is possible that *any two images* would be sufficient. To test this, we *randomly* selected images for the training set; see the "random" line in Figure 3. Each point in this line is the average over 10 random choices. We see that this line is well below the LMU line. Moreover, the segmenter remains statistically inferior to the "all" line until observing (on average) 5 randomly-drawn images.

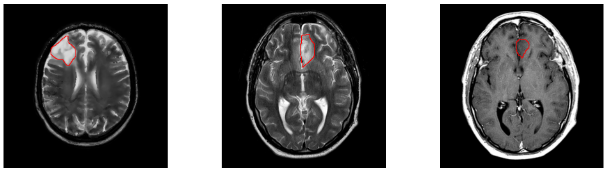

Figure 4. Sample images from brain tumor dataset, tumor is segmented in red

4.2. Finding Tumors in Brain Scans

Here, we consider the challenge of finding tumors within a patient's brain — that is, labeling each pixel in a magnetic resonance (MR) image as either tumorous or non-tumorous. This task is crucial in surgical planning and radiation therapy, and currently requires a significant amount of manual work by human medical experts.

Here, we have 80 images (axial slices) from the brains of 16 patients. These are taken from different regions of the brains; in particular, no two images are adjacent to each other. We resized each image from 256×256 pixels to 65×65. Figure 4 shows three images from this dataset, with the tumor segmentation outlined in red. We again segment these images into an unlabeled set U, containing 71 images from 11 individuals, and a test set T, containing 9 images from the other 5 patients.[3]

Features: Most patient visits yield scans in 3 different MR modalities: T1, T2, and T1c (that is, T1 after the patient has received a contrasting agent). We identify each pixel i with a vector of 4 values, including the T2 value and the difference between T1c and T1. As each brain is somewhat symmetric around the sagittal plane, we also include the symmetry feature by computing the difference between intensities of pairs of symmetrical pixels with respect to the sagittal plane, for both T2 and T1c - T1 modalities. So we compute four features for each pixel.

Figure 5 shows the results of actively selecting the training set. Actively training on one image in this dataset produces an average F-measure of 57%. This segmenter is as good as the one obtained using all of the data (paired t-test $p < 0.05$). (The specific image selected is the left one in Figure 4.) Training on the second LMU-selected image increased the average accuracy to 70%, which is higher than the 61% accuracy obtained using all of the data. (Note, however, that this is not statistically better, at $p < 0.05$.) Donmez and Carbonell (Donmez & Carbonell, 2008) report a similar situation (albeit in active sampling in rank learning on TREC 2004 dataset), noting that the performance of their active learning algorithm is sometimes better than the one obtained by training on all the data. Finally, as with the Sky data, on average the segmenters produced using the first several (here three) *randomly* drawn images were all significantly inferior to the "all" segmenter.

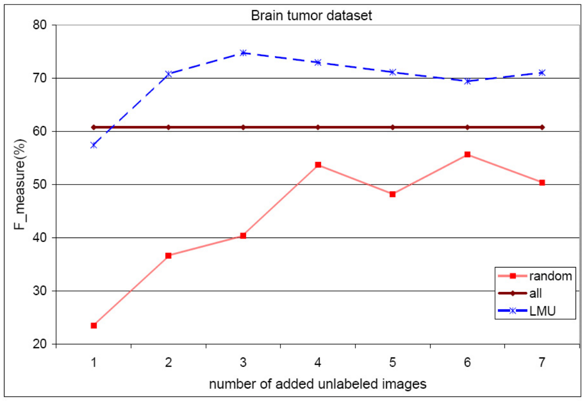

Figure 5. Accuracy of segmentation versus number of training data chosen from unlabeled set for brain tumor dataset

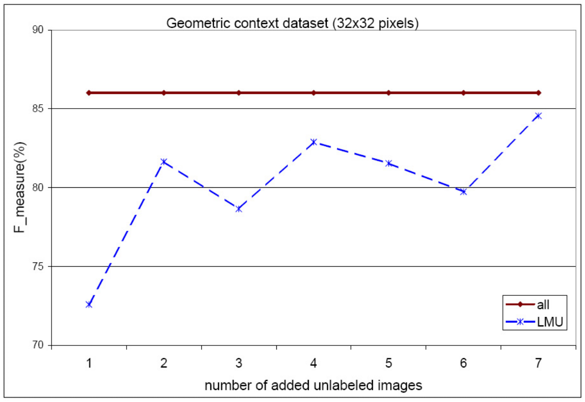

Figure 6. Accuracy of segmentation versus number of training data chosen from unlabeled set for geometric context dataset with images downsized to 32×32 pixels

4.3. Scalability Study

This subsection explores how our LMU scales with the size of the images. We therefore downsized the images in the geometric context dataset from 65×65 to 32×32, then repeated the same active learning process described above. The results, appearing in Figure 6, show essentially the same trend that appeared in Figure 3. Here, however, LMU required 7 images before it first obtained a segmenter whose performance was statistically "equivalent" to the one based on all of the data (paired t-test, $p < 0.05$).

[3] The F-measure score is problematic if any test image has no tumor, as here there can be no true positives ($tp = 0$). Hence, to simplify our analysis, our T includes only images that contain some tumor. However, the unlabeled set U includes some images that have no tumor.

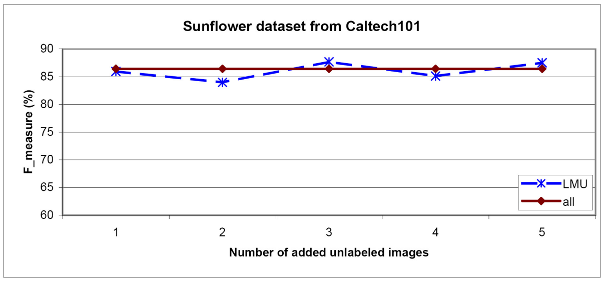

Figure 7. LMU of sunflower category in Caltech101 dataset

4.4. Discussion

It is, at first, very surprising that one can produce an effective segmenter with so few images — here, only 2 for the (original) sky data, and 1 for the brain tumor data! Towards explaining this, note that each of these images is not really a single "point", but is actually $65 \times 65 \approx$ 4,000 pixels, and each oracle-label is actually providing around 4000 bits. Hence, the 2 sky images is essentially 8000 bits, which is a lot of information. The results in the scaling studies are consistent with this conjecture: Here, we required 7 carefully-selected images to obtain the information needed to do well; notice that this $7 \times (32 \times 32) \approx 2 \times (65 \times 65)$.[4]

As further support, consider the segmenters based on randomly-drawn images. While they were, typically, worse than ones based on images specified by LMU, we found that we were still getting good segmenters using only a few such random images. Again, this is consistent with the view that the label of each image is supplying a great deal of information — even if the image is drawn randomly.

4.5. Other Results

The results shown above strongly suggest that very few images are sufficient to train a DRF-based segmentor, but only if they are well selected. We explored this claim over other datasets. Figure 7 shows the results of applying LMU on the set of sunflower objects from Caltech101 dataset (Fei-Fei et al., 2007). Since the dataset is intended for object recognition task, it is relatively easy to segment the object from background, which is probably why LMU gets good accuracy after actively selecting only one image.

We also explored many ways to reduce LMU's computation complexity. For example, the LMU-LR system used logical regression to estimate the conditional probabilities (as if the pixels were independent) for the entropy function. Note this is just used to select the appropriate image to give to the oracle; the oracle then finds the parameters for the full DRF, based on equation 6. (Greiner et al., 2009) presents those results.

We also experimented using several variants of our LMU, including one that used only RU throughout, and another that used only(a variant) of MU. (Greiner et al., 2009) presents those findings, which demonstrate that our LMU is superior.

5. Conclusions

While there are now many results in active learning, this is one of the first studies that considers the challenge of actively learning the parameters of a DRF-based segmenter. While our LMU system is based on standard "parts" — selecting the image with maximal uncertainty, and or that most reduces the uncertainty of other images — we found that this particular combination was effective for this task. We also found, to our surprise, that we could produce an effective segmenter using very few segmented images. Our studies support the claim that it may be because the label for a single image contains a great deal of information; *i.e.,* corresponds to receiving many single-bit labels, in the standard active learning framework.

Acknowledgments

All authors gratefully acknowledge the support from the Alberta Ingenuity Centre for Machine Learning (AICML), NSERC,and iCORE. We also thank Dr. Albert Murtha, as well as the Alberta Cancer Board, for providing the (segmented) brain tumor images. Csaba Szepesvári is on leave from MTA SZTAKI, Bp. Hungary.

References

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of Royal Statistical Society, Series B*, *48*, 259–302.

Campbell, C., Cristianini, N., & Smola, A. (2000). Query learning with large margin classifiers. *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 111–118).

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Collins, B., Deng, J., Li, K., & Fei-Fei, L. (2008). Towards scalable dataset construction: An active learning approach. *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 86–98).

[4] We are not claiming that 8000 bits is a magical number — instead, we are just observing that our active learner requires more small images than large images, which is consistent with the claim that the number of pixels being labeled seems significant.

Donmez, P., & Carbonell, J. G. (2008). Optimizing estimated loss reduction for active sampling in rank learning. *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 248–256).

Fei-Fei, L., Fergus, R., & Perona, P. (2007). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *106*, 59–70.

Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, *28*, 133–168.

Greiner, R., Farhangfar, A., & Szepesvari, C. (2009). http://www.cs.ualberta.ca/∼greiner/RESEARCH/ActiveLearning4Segmentation/.

Guo, Y., & Greiner, R. (2007). Optimistic active learning using mutual information. *International Joint Conference on Artificial Intelligence(IJCAI)* (pp. 823–829).

Kumar, S., & Hebert, M. (2003). Discriminative random fields: A discriminative framework for contextual interaction in classification. *Proceedings of the International Conference on Computer Vision (ICCV)* (pp. 1150–1157).

Lee, C., Wang, S., Brown, M., Murtha, A., & Greiner, R. (2008). Segmenting brain tumors using pseudo-conditional random fields. *Proceedings of the 11th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* (pp. 359–366).

Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. *Proceedings of the Special Interest Group on Information Retrieval (ACM-SIGIR)* (pp. 13–19).

Nguyen, H., & Smeulders, A. (2004). Active learning using pre-clustering. *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 623–630).

Roy, N., & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 441–448).

Schohn, G., & Cohn, D. (2000). Less is more: Active learning with support vector machines. *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 839–846).

Szummer, M., Kohli, P., & Hoiem, D. (2008). Learning CRFs using graph cuts. *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 582–595).

Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*, 45–66.

Varma, M., & Zisserman, A. (2002). Classifying images of materials: Achieving viewpoint and illumination independence. *Proceedings of the 7th European Conference on Computer Vision(ECCV)* (pp. 255–271).

Vijayanarasimhan, S., & Grauman, K. (2008). Multi-level active prediction of useful image annotations for recognition. *Proceedings of the Neural Information Processing Systems (NIPS)* (pp. 1705–1712).

Xu, Z., Yu, K., Tresp, V., Xu, X., & Wang, J. (2003). Representative sampling for text classification using support vector machines. *Proceedings of the European Conference on Information Retrieval* (pp. 393–407).

Zhang, C., & Chen, T. (2002). An active learning framework for content based information retrieval. *IEEE trans. on Multimedia, Special Issue on Multimedia Database*, *4*, 260–268.