Alter-CNN: An Approach to Learning from Label Proportions with Application to Ice-Water Classification

Fan Li
University of Guelph
Guelph, ON N1G 2W1
fli02@uoguelph.ca

Graham Taylor
University of Guelph
Guelph, ON N1G 2W1
gwtaylor@uoguelph.ca

Abstract

We present an approach to train a model for classifying ice and open water directly using the polygon-wise ice concentration available from ice charts. This can be considered as a "learning from label proportions" (LFLP) problem which has been studied in the last decade and applied to many real-world applications. Our approach is based on convolutional neural networks (CNNs), which have been shown to capture representative features and achieve impressive classification results provided a large number of *labeled* training samples. We provide a probabilistic formulation to learn from label proportions while considering the proportion bias, and an expectation-maximization (EM) approach is employed to estimate both the CNN model parameters and infer the per-pixel labels. Experiments on a large-scale satellite image dataset[1] show that our proposed approach achieves better results than previous approaches for LFLP problems.

1 Introduction

Operational sea ice mapping is very important for ship navigation, environmental study, and many other purposes. Currently, mapping of sea ice charts is performed daily by a team of experts from the Canadian Ice Service. These ice charts are formed largely based on their visual interpretation of satellite imagery. Due to the massive scale of the sea area, accurate high-resolution mapping (e.g. at the pixel level) is not practical. Instead, the ice experts first divide the sea area into multiple polygons, and then estimate the ice concentration, ice types, and other information in each polygon. An example of ice charts is shown in Figure 1, in which the ellipses are called "egg codes" defined by the World Meteorological Organization.

Even though current ice charts provide helpful information about ice conditions at a macroscopic level, they lack detail because the number of egg code polygons is limited. Mapping requires the effort and experience of the ice experts, and there might also exist inter-operator bias among different ice experts [2]. Therefore, classification algorithms for distinguishing ice and open water have the potential to provide higher resolution, more accurate, and more systematic ice charts.

Previously, supervised classification methods such as neural networks [3] and support vector machines (SVM) [4] have been successfully applied for sea ice classification. However, the delineation of pixelwise ground truth is difficult and time-consuming due to the large image size, the complexity of ice and weather conditions, and the uncertainty in some transiting areas. If only a small number of pixels are used for training, they might be incapable of representing the data distribution due to the within-scene and across-scene variation of ice and water characteristics.

Considering the large amount of ice charts produced in recent years, developing an algorithm that can use the ice charts directly for training would very useful. Learning a model for ice-water classification using the ice concentration, i.e., the proportion of the ice class in the polygons, can be considered as a LFLP problem by treating the polygons as bags. This topic has been studied recently in the machine learning community, and mainly

[1] http://hdl.handle.net/10864/WJ7VY

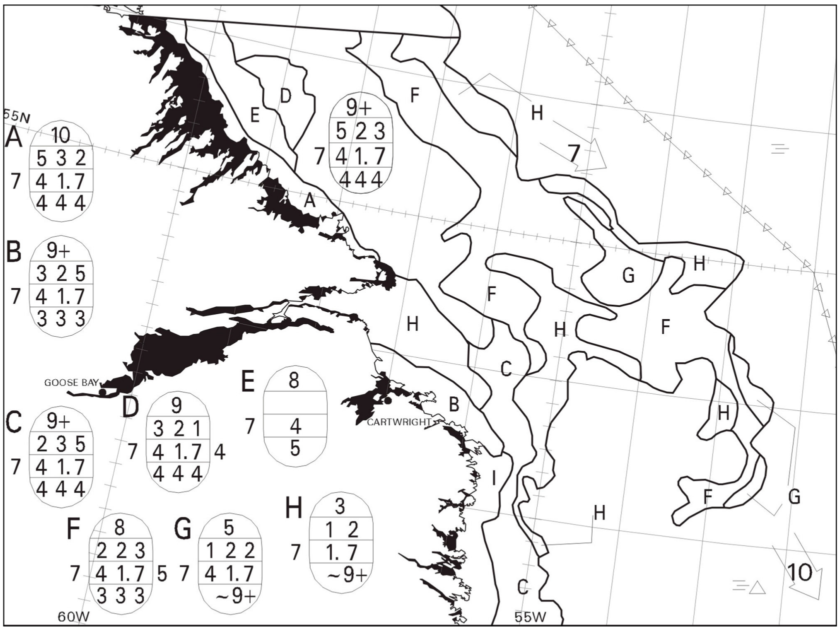

Figure 1: An ice chart [1] with the egg codes, recording the total ice concentration, the partial ice concentration, the stage of ice development, and the form of ice in the polygon. For example, the egg code G means that the polygon consists of 50% ice, including 10% gray ice, 20% medium first-year ice, and 20% first-year thin ice.

applied to situations where there are privacy constraints or when the labels are expensive to collect [5, 6, 7]. Our problem and many other remote sensing classification problems fall under the second scenario. The main difficulty of our problem compared to those related to privacy constraints is that the ice concentration is estimated by human beings instead of being calculated from the true labels, and thus may be biased from the true proportions.

In this paper, we propose a method based on CNNs and probabilistic graphical models to learn a classifier from label proportions. In recent years, CNNs have been widely used for a variety of computer vision applications. It has been found that CNNs are capable of representing the data better than hand-crafted feature descriptors for natural images [8], medical images [9], and recently for remote sensing images [10]. However, the success of CNNs has mirrored and exploited the growth of large labeled datasets, such as ImageNet [8]. Our work departs from previous approaches that employ CNNs to output per-pixel predictions, such as scene parsing [11, 12], in that dense pixelwise labels are not available. The proposed algorithm is evaluated using both the label proportions derived from per-pixel ground truth at different noise levels and those provided by the ice experts.

2 Related work

The LFLP problem is a special case of weakly supervised learning, and is also related to other types of learning methods. If there is only one class in each bag, this problem is reduced to standard classification. If the proportions are unknown, it becomes a multiple instance learning problem. Class uncertainty is also associated with the label proportions. The most uncertain case is when the class proportions are uniform across a bag. Fortunately, this situation is avoided by ice experts when they form the polygons, and this will be discussed later in Section 5.1.

Despite the wide-applicability of the LFLP problem, it received little attention up until the last decade. Kück and de Freitas [13] first addressed the problem by building a generative model and using the MCMC sampler for training. Quadrianto et al. [5] provided a formal description of this problem, and proposed a mean operator to reconstruct the labels that could offer the same performance guarantees of uniform convergence as the fully-supervised counterparts. However, their method is based on the assumption that the conditional distribution of the data is independent of the bags when the labels are given. This assumption is not satisfied when the samples are not grouped into bags by random. For the ice-water classification problem, for instance, the data distribution is highly dependent on the bags.

More recently, Rueping [6] proposed a method that applies a SVM at the bag level. The mean of the bags and their labels are calculated from the label proportions using an inverse calibration method. Yu et al. [7] also formulated the problem within an SVM framework, and attempted to optimize the labels and the model parameters simultaneously. To address the non-convex integer programming problem, they proposed two algorithms to solve the problem, the first one using an alternating optimization method and the second one using an convex relaxation method. Our method is very close to their first method. However, our gradient-based formulation permits a much larger and more complex family of models that have the capability of learning rich feature representations: so-called deep neural networks.

Recently, several approaches have applied CNNs in the weakly supervised learning setting for computer vision tasks. Oquab et al. [14] found that by only training on image-level tags, CNNs can generate accurate image-level labels and predict approximate locations of objects, performing comparably to fully-supervised approaches. Pathak et al. [15] proposed a method to learn pixelwise labeling from image-level tags by imposing linear constraints on the output labeling of a CNN classifier. Kotzias et al. [16] proposed an objective function that leverages instance-based similarity and group-level label information based on deep learning methods to learn instance-based classifiers. Papandreou et al. [17] provided a probabilistic formulation for semantic image segmentation based on CNN under both weakly-supervised and semi-supervised settings, and solved it by using an EM method. Our approach can be considered as an extension of their approach to the LFLP problem.

3 Preliminaries

For a LFLP problem, we aim to learn a classifier to predict the pixelwise labels $\mathbf{y}$ given the image patches $\mathbf{x}$, the bag information $\mathcal{B}$, and the label proportion $\mathbf{z}$ for each bag. Each bag $\mathcal{B}_i$ contains a set of image patches $\mathcal{S}_i$. $|\mathcal{B}|$ is the number of bags, and $|\mathcal{S}_i|$ is the number of patches in $\mathcal{B}_i$. Our specific problem is a binary classification problem, i.e., $\mathbf{y} \in \{-1, 1\}$, and we define +1 for ice and -1 for open water. The label proportion $\mathbf{z} \in [0, 1]$ is the proportion of the ice in a bag. If the label proportion is accurate, it can be represented by the labels in the bag: $z_i = \frac{\sum_{s=1}^{|\mathcal{S}_i|} y_s}{2|\mathcal{S}_i|} + \frac{1}{2}$. In practice, however, there is a bias between the label proportion estimated by an ice expert and the true proportion, no matter how experienced the ice expert is. In Section 4, a probabilistic framework is presented to model this LFLP problem by considering the proportion bias.

4 The proposed algorithm

The LFLP problem can be formulated into a probabilistic graphical model:

$$P(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta) = P(\mathbf{x}) \prod_{i=1}^{|\mathcal{B}|} \{ \prod_{s=1}^{|\mathcal{S}_i|} P(y_s \mid \mathbf{x}; \theta) P(z_i \mid \mathbf{y_i}) \} \tag{1}$$

where θ are the CNN parameters.

In our problem, the image data $\mathbf{x}$ and the label proportion of each bag z_i are observed, and the pixelwise labels y_s are latent variables. We make the assumption that the label proportions are independent of the image data if the labels are given. $P(\mathbf{z} \mid \mathbf{y})$ can thus be defined as

$$P(z_i \mid \mathbf{y_i}) = \frac{1}{Z} \exp\{-|\hat{z}_i - z_i|\} = \frac{1}{Z} \exp\left\{ -\left| \frac{\sum_{s=1}^{|\mathcal{S}_i|} y_s}{2|\mathcal{S}_i|} + \frac{1}{2} - z_i \right| \right\} \tag{2}$$

where $\hat{\mathbf{z}}$ are the label proportions calculated from the predicted labels, and $Z = \int_0^1 \exp\{ - |\hat{z}_i - z_i| \} \, \mathrm{d}z_i$ is a normalization constant.

An EM approach is used to alternatively infer the pixelwise labels $\mathbf{y}$ and update the model parameters θ from the training data similar to [17]. In the E step, the labels are estimated by

$$\begin{aligned}
\mathbf{y} &= \underset{\mathbf{y}}{\operatorname{argmax}}\, P(\mathbf{y} \mid \mathbf{x}; \theta') P(\mathbf{z} \mid \mathbf{y}) \\
&= \underset{\mathbf{y}}{\operatorname{argmax}}\, \{\log P(\mathbf{y} \mid \mathbf{x}; \theta') + \log P(\mathbf{z} \mid \mathbf{y})\} \\
&= \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{i=1}^{|\mathcal{B}|} \left\{ \sum_{s=1}^{|\mathcal{S}_i|} f_s(y_s \mid \mathbf{x}; \theta') - |\mathcal{S}_i| \left| \frac{\sum_{s=1}^{|\mathcal{S}_i|} y_s}{2|\mathcal{S}_i|} + \frac{1}{2} - z_i \right| \right\}
\end{aligned} \tag{3}$$

where $f_s(y_s \mid \mathbf{x}; \theta')$ is the output of the CNN classifier for sample s.

From (3), we can see that instead of fixing the labels entirely by the label proportions provided, the optimum label configuration is determined by leveraging the predictions made by the CNN classifier and the observed label proportion. This is important for our application due to the existence of the proportion bias.

To solve (3), we can use the bag-wise optimization method in [7] which is also designed for a binary classification problem. For each bag $\mathcal{B}_i$, we first initialize all the labels into -1, and sort the samples based on the value of $f_s(y_s \mid \mathbf{x}; \theta')$. Then, labels are flipped from the one with smallest value. After each flip, the log-likelihood for the corresponding label configuration and label proportion is calculated. [7] has shown that the maximum log-likelihood value by incrementally flipping all the labels is the optimal solution.

In the M step, the complete data log-likelihood $Q(\theta; \theta') \approx \log P(\hat{y} \mid \mathbf{x}; \theta)$ is optimized by mini-batch stochastic gradient descent (SGD) with fixed $\mathbf{y}$ similar to [17]. The complete algorithm is shown in Algorithm 1.

Algorithm 1 The Alter-CNN algorithm for binary classification from label proportions.

```
Inputs:
    Image data $\mathbf{x}$, bags $\mathcal{B}$, label proportions of each bag $z_i$, $i \in \{1, ..., |\mathcal{B}|\}$, initial CNN
    parameters $\theta'$, and initial labels $\mathbf{y}$.
E-step: Optimize $\mathbf{y}$. For each bag $\mathcal{B}_i$:
    Initialize all of the labels $y_i$ to -1.
    Sort the samples based on $f_s(y_s \mid \mathbf{x}; \theta')$.
    Flip the negative labels incrementally from the smallest $f_s(y_s \mid \mathbf{x}; \theta')$ value, and calculate the correspond-
    ing log-likelihood using (3).
    Change the labels of bag $\mathcal{B}_i$ to the label configuration corresponding to the maximum log-likelihood value.
M-step: Optimize $\theta$.
    $Q(\theta; \theta') = \log P(\hat{\mathbf{y}} \mid \mathbf{x}, \theta) = \sum_{i=1}^{|\mathcal{B}|} \sum_{s=1}^{|\mathcal{S}_i|} \log P(\hat{y_s} \mid \mathbf{x}, \theta)$.
    Calculate $\bigtriangledown_\theta Q(\theta; \theta')$ and use SGD to update $\theta'$.
```

5 Experiments

We demonstrate the performance of our approach using a sea ice dataset. The proposed Alter-CNN algorithm is compared with Invcal-SVM [6] and Alter-$\propto$SVM [7] on hand-crafted features that have been demonstrated to work well for similar data and classification tasks. The Alter-CNN is implemented using Theano. The tests are performed on a workstation with 2 Intel Xeon E5-2620 CPUs (6-core each and 2.10GHz), and a Nvidia Titan Black GPU.

5.1 Dataset

The dataset used in our experiments is obtained from the C-band RADARSAT-2 SAR satellite over the Beaufort and Chukchi Sea area from May to December in the year 2010. Compared to other sensors, the SAR sensors are capable of all-weather and all-day imaging which is important for sea ice monitoring. The data were captured in the ScanSAR Wide mode, which is the most useful beam mode for sea-ice monitoring. HH and HV dual-polarizations are provided in the ScanSAR Wide mode. The image sizes are around 2,500 × 2,500 pixels after performing 4 × 4 block averaging on the original images. The spatial resolution of all these images is 50 metres. The human estimate of the ice concentration of each egg code polygon for all the images is provided by an experienced ice analyst, and accurate pixelwise ground truth for each image has been created by considering both the images and the estimated ice concentration. More information about the dataset can be found in [4].

In our experiments, we exclude images which contain only ice or only open water in order to better evaluate the performance of the LFLP methods, though including those images can further improve classification performance. Ten high-resolution images containing 140 bags are available after excluding the homogeneous images. The correct proportion of ice calculated from the pixelwise ground truth and the estimated proportion by ice experts are compared in Figure 2. The ice experts usually try to reduce label ambiguity by drawing polygons that contain a majority of ice or open water, and therefore the proportions mainly lie at opposite ends of the histogram. The total percentage of ice for all the images is 51.68%, which is very close to the estimated percentage by ice experts (52.19%). However, the average squared error of the estimated proportions weighted by the number of samples in the bags is 85.39%, which verifies that macroscopically the ice experts are good at estimating the ice concentration, but for a specific area they may make mistakes and there may be inconsistency among the different ice experts.

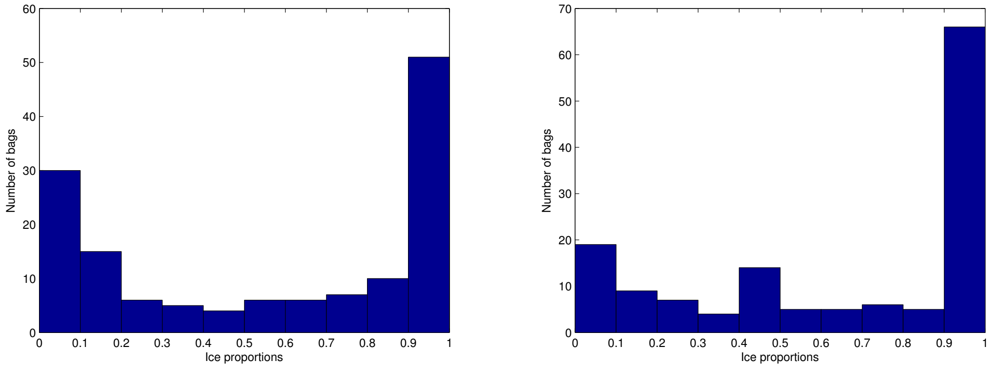

Figure 2: The histogram of the correct ice proportions (left) and the estimated proportions by ice experts (right) for all of the bags.

5.2 Separation of training and test data

Since some of the images have very unique data distributions, the classification performance of image-wise training and testing is poor even if the pixelwise labels are used. Therefore, we perform bag-wise training and testing. A five-fold cross-validation is performed by randomly selecting bags for all the images. For the CNN, a total number of 10,000 patch samples are selected from each image for training and testing, and 20 percent of the bags are further separated from the training set for validation. For Alter-$\propto$SVM and Alter-CNN, 3000 samples are selected per image due to the computational costs of processing large kernel matrices. We verified that there is no difference in terms of cross-validation accuracy using the number of samples per image ranging from 500 to 5000.

5.3 Initialization

Initialization is important for Alter-$\propto$SVM and Alter-CNN because they are both based on an alternating optimization. In [7], random sampling is used for the initial labeling, and the best result is chosen from the one with lowest objective function in multiple tests. This is very time-consuming, and good solutions cannot be be easily found when the training sample size is huge. Earlier we have shown that the bag proportions are mainly close to 0% or close to 100%. Here we use a stochastic method for Alter-$\propto$SVM and Alter-CNN. Each label y_s in bag $\mathcal{B}_i$ is initialized stochastically based on the label proportion z_i of the bag: $P(y_s = +1) = z_i$. For Alter-CNN, the model is trained on the stochastic labels for the first a few iterations before using EM.

5.4 CNN architectures

We use a multi-layer CNN architecture similar to the LeNet-5 [18] designed for handwritten digit classification. The input image patches are 21 × 21, and there are a total of 882 inputs considering the dual-polarization bands. In total, the network has six layers, including two convolutional layers, two max-pooling layers, a fully-connected layer, and a fully-connected output layer. The first convolutional layer has 20 feature maps, and the second convolutional layer has 50 feature maps, both of which use 4 × 4 filters. The hyperbolic tangent activation function is used for the hidden layers. The number of output neurons in this layer is set to 100. The logistic unit is used for the output layer. The batch size for SGD is 500.

5.5 CNN training

Because the labels of the validation data are not available during the training step, the label proportion of the validation data is used directly for validation. We use the weighted squared error calculated by

$$\epsilon = \sqrt{\frac{\sum_{i=1}^{|\mathcal{B}|}(z_i - \hat{z}_i)^2|\mathcal{S}_i|}{\sum_{i=1}^{|\mathcal{B}|}|\mathcal{S}_i|}} \tag{4}$$

Due to the loss of information using only label proportions as well as the proportion bias, the reduction of the proportion error does not always reflect a better solution. However, experimental results show that the proportion error of the validation data can still be used as a measure of model performance in a certain level. Early stopping is used based on the validation error. The maximum number of epoch iterations is set to 100, including the first 20 iterations training on the stochastic labels. Also, optimizing the labels after performing SGD for each epoch is unnecessary. In our experiment, the re-labeling is performed only when the validation error is reduced in order to prevent assigning the labels based on a poor solution.

5.6 Hand-crafted features

For Invcal-SVM and Alter-∝SVM, we use two sets of features for comparison. The grey-level co-occurrence matrix (GLCM) features [19] are so far among the best texture features for SAR sea ice classification [20, 4]. The first feature set contains both the averaged intensity and GLCM textures including entropy, contrast, and correlation, using the same image patches for the CNN. The second feature set contains the mean, standard deviation, and GLCM measures in different window sizes, which are selected from the same dataset using the pixelwise ground truth and the forward selection method based on the SVM. These features have been used in [4] for fully-supervised ice-water classification. Using the second feature set will give an unfair advantage over our approach because in practice, the pixelwise ground truth for the dataset is not available. There are in total 8 features in the first set and 28 features in the second set.

5.7 Experimental results

We test the algorithms using the correct label proportions calculated from the pixel-wise ground truth and the ice concentration estimated by the ice experts. We also add white Gaussian noise with standard deviations ranging from 0.1 to 0.3 to the correct proportions to evaluate the robustness of the classifiers under noisy labeling conditions. The linear kernel is used for Invcal-SVM and Alter-∝SVM.

The results are shown in Table 1. For the tests with Gaussian noise, the tests are repeated five times and the averaged classification accuracy is reported. We see that the Invcal-SVM approach is evidently worse than the other methods. This indicates that the bag mean is incapable of representing the properties of all the samples in the bag. Alter-∝SVM using the hand-crafted features selected from the same dataset and the pixelwise ground truth achieves significantly better results than Invcal-SVM and the same method using features with a fixed window size. Alter-CNN achieves the best classification accuracy for all the cases even with a fixed window size. Also, both Alter-∝SVM and Alter-CNN are robust to the label proportion bias because the optimization of the labels is based on both the classifier and the given label proportion. The classification accuracy has very little decrease when the standard deviation is within 20%. Using the estimated proportions with an average error of 85%, 87.75% cross-validation accuracy can still be achieved by Alter-CNN.

6 Conclusions

This paper proposed a CNN-based approach to learn a model to classify ice and open water directly using label proportions. This weak label information is provided by ice charts made by ice experts. We formulated the problem as a probabilistic graphical model, and applied EM to alternate between updating the latent pixelwise labels and the CNN parameters. Results on a sea ice dataset show that our method outperforms previous methods trained

Table 1: Five-fold cross-validation accuracy (%) of Invcal-SVM and Alter-$\propto$SVM using different features, and Alter-CNN.

Methods (features used)	Invcal-SVM (8)	Invcal-SVM (28)	Alter-$\propto$SVM (8)	Alter-$\propto$SVM (28)	Alter-CNN (882)
Correct proportions	56.78	60.26	76.03	85.82	**89.50**
Added ± 0.1 noise	61.57	61.08	76.31	85.21	**87.23**
Added ±0.2 noise	59.61	60.18	76.12	85.31	**87.84**
Added ±0.3 noise	57.96	58.94	76.72	82.83	**83.57**
Estimated proportions	49.83	50.66	76.36	85.53	**87.75**

on well-selected hand-crafted features using both simulated proportions in different noise levels and the estimated proportions by the ice experts. Our method also has the potential to be used for other remote sensing classification tasks in which the pixelwise ground truth is difficult to obtain. Potential future work is the multi-class extension of our approach that can be applied to the classification of multiple sea ice types.

Acknowledgments

The authors would like to thank MacDonald, Dettwiler and Associates for the RADARSAT-2 dataset and ice analyst D. Isaacs for providing the corresponding ground truth. RADARSAT-2 data and products are copyright of MacDonald, Dettwiler and Associates Ltd. 2011, all rights reserved.

References

[1] D. Fequest, *MANICE: manual of standard procedures for observing and reporting ice conditions*. Environment Canada, 2005.

[2] M.-A. Moen, A. P. Doulgeris, S. N. Anfinsen, A. H. Renner, N. Hughes, S. Gerland, and T. Eltoft, "Comparison of feature based segmentation of full polarimetric SAR satellite sea ice images with manually drawn ice charts," *The Cryosphere*, vol. 7, no. 6, pp. 1693–1705, 2013.

[3] J. A. Karvonen, "Baltic sea ice SAR segmentation and classification using modified pulse-coupled neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1566–1574, 2004.

[4] S. Leigh, Z. Wang, D. Clausi *et al.*, "Automated ice-water classification using dual polarization SAR satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5529–5539, 2014.

[5] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, "Estimating labels from label proportions," *J. Mach. Learn. Res*, vol. 10, pp. 2349–2374, 2009.

[6] S. Rüping, "SVM classifier estimation from group probabilities," in *ICML*, 2010, pp. 911–918.

[7] F. Yu, D. Liu, S. Kumar, T. Jebara, and S. Chang, "$\propto$SVM for learning with label proportions," in *ICML*, 2013.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[9] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *NIPS*, 2012, pp. 2843–2851.

[10] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "DeepSat-A learning framework for satellite imagery," *arXiv preprint arXiv:1509.03602*, 2015.

[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Scene parsing with multiscale feature learning, purity trees, and optimal covers," in *ICML*, 2012.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *ICLR*, 2015.

[13] H. Kuck and N. de Freitas, "Learning about individuals from group statistics," in *UAI*, 2005.

[14] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?–weakly-supervised learning with convolutional neural networks," in *CVPR*, 2015, pp. 685–694.

[15] D. Pathak, P. Krähenbühl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," *arXiv preprint arXiv:1506.03648*, 2015.

[16] D. Kotzias, M. Denil, N. De Freitas, and P. Smyth, "From group to individual labels using deep features," in *SIGKDD*, 2015, pp. 597–606.

[17] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a DCNN for semantic image segmentation," in *ICCV*, 2015.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Trans. Sys. Man Cyber.*, no. 6, pp. 610–621, 1973.

[20] D. A. Clausi and B. Yue, “Comparing cooccurrence probabilities and Markov random fields for texture analysis of SAR sea ice imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 215–228, 2004.