Joint Calibration for Semantic Segmentation

Holger Caesar
holger.caesar@ed.ac.uk
Jasper Uijlings
jrr.uijlings@ed.ac.uk
Vittorio Ferrari
vittorio.ferrari@ed.ac.uk

School of Informatics
University of Edinburgh
Edinburgh, UK

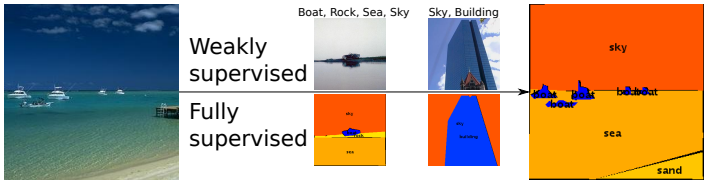

Figure 1: Semantic segmentation is the task of assigning class labels to all pixels in the image. Depending on the setting, we have either pixel-level or image-level ground-truth labels available at training time.

Figure 2: Fully supervised semantic segmentation on SIFT Flow. We show an example image, the target output, our uncalibrated output and our jointly calibrated output.

Semantic segmentation is the task of assigning a class label to each pixel in an image (Fig. 1). In the fully supervised setting, we have ground-truth labels for all pixels in the training images. In the weakly supervised setting, class-labels are only given at the image-level. We tackle both settings in a single framework which builds on region-based classification.

Our framework addresses three problems common to region-based semantic segmentation. First of all, objects naturally occur at different scales within an image [3]. Performing recognition at a single scale inevitably leads to regions covering only parts of an object which may have ambiguous appearance, and to regions straddling over multiple objects, whose classification is harder due to their mixed appearance. Therefore many recent methods operate on pools of regions computed at multiple scales, which have a much better chance of containing some regions covering complete objects [1, 2]. However, this leads to overlapping regions which may lead to conflicting class predictions at the pixel-level. These conflicts need to be properly resolved.

Secondly, classes are often unbalanced [2, 4]: "cars" and "grass" are frequently found in images while "tricycles" and "gravel" are much rarer. Due to the nature of most classifiers, without careful consideration these rare classes are largely ignored: even if the class occurs in an image the system will rarely predict it. Since class-frequencies typically follow a power-law distribution, this problem becomes increasingly important with the modern trend towards larger datasets with more and more classes.

Finally, classes compete: a pixel can only be assigned to a single class (e.g. it can not belong to both "sky" and "airplane"). To properly resolve such competition, a semantic segmentation framework should take into account predictions for multiple classes jointly.

In this paper we address these three problems with a joint calibration method over a set of SVMs.

Model. We represent an image by a set of overlapping regions [3] described by CNN features [1]. Our semantic segmentation model infers the label o_p of each pixel p in an image:

$$o_p = \underset{c,\, r \ni p}{\arg\max}\ \sigma(w_c \cdot x_r,\, a_c, b_c) \qquad (1)$$

As appearance models, we have a separate linear SVM w_c per class c. These SVMs score the features x_r of each region r. The scores are calibrated by a sigmoid function σ, with different parameters a_c, b_c for each class c. The arg max returns the class c with the highest score over all regions that contain pixel p. This involves maximizing over classes for a region, and over the regions that contain p.

Joint Calibration. The final goal of semantic segmentation is to output a pixel-level labeling, which is evaluated in terms of pixel-level accuracy. Now, using uncalibrated SVMs is problematic. SVMs are trained to predict class labels at the region-level, not the pixel-level. However, different regions have different area, and, most importantly, not all regions contribute all of their area to the final pixel-level labeling: Predictions of small regions may be completely suppressed by bigger regions. In other cases, bigger regions may be *partially* overwritten by smaller regions. Furthermore, the SVMs are trained in a one-vs-all manner and are unaware of the competition between classes. To address these problems we propose to jointly calibrate SVMs for the final evaluation measure. We do this by applying sigmoid functions σ to all SVM outputs. We calibrate the parameters a_c, b_c for all classes jointly by minimizing a loss function that depends on the pixel labeling output of our method and the ground-truth labeling. Since we optimize for the accuracy of this final output labeling, and we do so jointly over classes, our calibration procedure takes into account both problems of conflicting class predictions between overlapping regions and competition between classes. Moreover, we also address the problem of class imbalance, as we compensate for it in our loss functions. We minimize our loss functions using coordinate descent. We iteratively apply line search to optimize the loss over a single parameter at a time, keeping all others fixed.

Experiments. We evaluate our method using class-average pixel accuracy (Cl. Acc.) on the challenging SIFT Flow dataset. Table 1 shows that we outperform the state-of-the-art for the default setting in semantic segmentation by 2.8% in the fully supervised and 2.4% in the weakly supervised setting. Fig. 2 shows an example of our method that is able to detect even small and rare objects in an image.

Method	Cl. Acc.
Tighe ECCV 2010	29.1%
Pinheiro ICML 2014	30.0%
Shuai CVPR 2015	39.7%
Tighe CVPR 2013	41.1%
Kekeç BMVC 2014	45.8%
Sharma NIPS 2014	48.0%
Yang CVPR 2014	48.7%
George CVPR 2015	50.1%
Farabet PAMI 2013	50.8%
Long CVPR 2015	51.7%
Sharma CVPR 2015	52.8%
Ours SVM	28.7%
Ours SVM+PS	27.7%
Ours SVM+JC	**55.6%**

Method	Cl. Acc.
Vezhnev. ICCV 2011	14.0%
Vezhnev. CVPR 2012	21.0%
Zhang TM 2014	27.7%
Xu CVPR 2014	27.9%
Zhang CVPR 2015	32.3%
Xu CVPR 2015	35.0%
Xu CVPR 2015 (transductive)	41.4%
Ours SVM	21.2%
Ours SVM+PS	16.8%
Ours SVM+JC	**37.4%**

Table 1: Class-average pixel accuracy in the fully supervised (left) and the weakly supervised setting (right) setting. We show results for our model on the test set of SIFT Flow using uncalibrated SVM scores (SVM), traditional Platt scaling (PS) and joint calibration (JC).

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[2] A. Sharma, O. Tuzel, and D. W. Jacobs. Deep hierarchical parsing for semantic segmentation. In *CVPR*, 2015.

[3] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[4] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *CVPR*, 2015.