Joint calibration of Ensemble of Exemplar SVMs

Davide Modolo[1], Alexander Vezhnevets[1], Olga Russakovsky[2], Vittorio Ferrari[1]
[1]University of Edinburgh [2]Stanford University.

The Ensemble of Exemplar SVMs [8] (EE-SVM) is a powerful non-parametric approach to object detection. It is widely used [1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13] because it explicitly associates a training example to each object it detects in a test image. This enables transferring meta-data such as segmentation masks [8, 12], 3D models [8], styles and viewpoints [1], GPS locations [6] and part-like patches [2]. Furthermore, EE-SVM can also be used for discovering objects parts [4, 10], scene classification [7, 10], object classification [3], image parsing [12], image matching [9], automatic image annotation [13] and 3D object detection [11].

An EE-SVM is a large collection of linear SVM classifiers, each trained from one positive example and many negative ones (an E-SVM). At test time each window is scored by all E-SVMs, and the highest score is assigned to the window. Because of this max operation, it is necessary to calibrate the E-SVMs to make their scores comparable. A common procedure is to calibrate each SVM independently, by fitting a logistic sigmoid to its output on a validation set [8]. Such independent calibration, however, does not take into account that the final score is the max over many E-SVMs. Moreover, calibrating one E-SVM in isolation requires choosing which positive training samples it should score high and which ones it can afford to score low. Such a prior association of positive training samples to E-SVMs is arbitrary, as there is no predefined notion of how much and in which way a particular E-SVM should generalize. What truly matters is the interplay between all E-SVMs through the max operation.

In this paper we present a *joint* calibration procedure that takes into account the max operation. We calibrate all E-SVMs at the same time by optimizing their joint performance *after* the max. Our method finds a threshold θ for each E-SVM, so that (i) all positive windows are scored positively by at least one E-SVM, and (ii) the number of negative windows scored positively by any E-SVM is minimized. The first criterion ensures that there are no positive windows scored negatively after the max, while the second criterion minimizes the number of false positives. We formalize these two criteria in a well-defined constrained optimization problem:

$$\min_{\Theta=\{\theta_j\}_{j=1}^{E}} \overbrace{\sum_{x\in\mathcal{N}} \mathbb{1}[\max_j(w_j\cdot x-\theta_j)]}^{\mathcal{L}(\Theta)} \quad (1)$$
$$s.t.\ \mathbb{1}[\max_j(w_j\cdot x-\theta_j)]>0,\ \forall x\in\mathcal{P}$$

The first requirement is formalised in its constraints, while the second comes in as a loss function to be minimized. In the equation, $\mathbb{1}$ is the indicator function and $\mathcal{P}$ and $\mathcal{N}$ are the sets of positive and negative windows in the training set. Furthermore, the ensemble contains E classifiers $\{w_j\}_{j=1}^{E}$. Each threshold θ_j defines which training samples the respective E-SVM e_j is scoring positively. By lowering a threshold we cover more positives and thereby satisfy more constraints, but we also include more negatives and therefore suffer a greater loss. Any positive sample can be potentially covered by any E-SVM, but at a different loss. We refer to a configuration Θ satisfying all the constraints as a *feasible solution*. Calibration is performed by adjusting the thresholds Θ.

The combinatorial nature of the problem makes it difficult to find the global optimum. We represent the space of all possible solutions as a search tree (fig. 1) and we propose an efficient, globally optimal optimization technique. By exploiting the structure of the problem we are able to identify areas of the solution space that cannot contain the optimal solution and discard them early on. Our globally optimal algorithm is able to calibrate a few hundred E-SVMs quickly. In order to solve larger problems with thousands of E-SVMs, we present a simple modification of our exact algorithm to deliver high quality approximate solutions.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

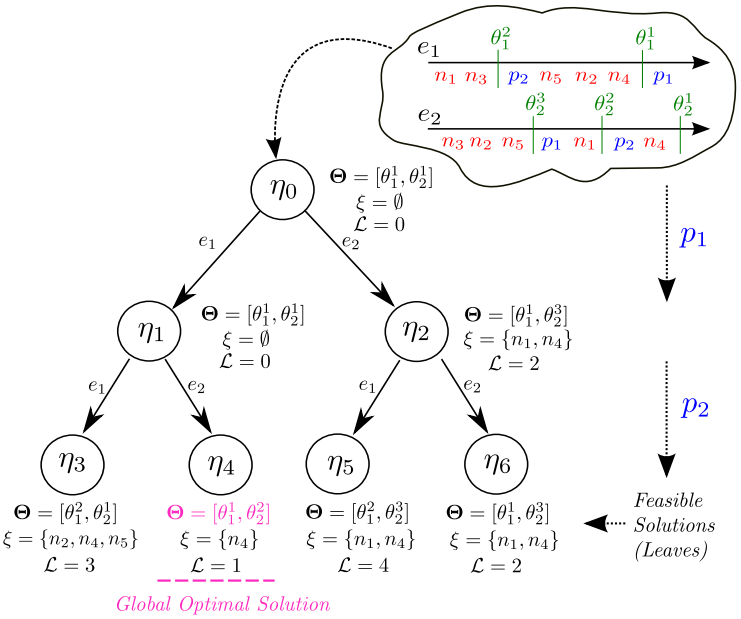

Figure 1: *Illustration of our joint calibration algorithm.* ***(Top)*** *the cloud shows (i) the initial scores given by two un-calibrated E-SVMs e_1 and e_2 on some training windows (positive and negative) and (ii) the E-SVMs candidate thresholds θ_j^e.* ***(Bottom)*** *the tree represents the space of all possible solutions. Θ is a configuration of E-SVMs thresholds and $\mathcal{L}$ is the loss function of our optimization problem. $\mathcal{L}$ counts the number of negative windows scored positively after the max operation ($|\xi|$). Note how the only feasible threshold configurations are those in the leaves (eq. 1).*

We train EE-SVM on state-of-the-art CNN descriptors [5] and we present experiments on 10 classes from the ILSVRC 2014 dataset and 20 from PASCAL VOC 2007. Our joint calibration procedure outperforms the classic independent sigmoid calibration [8] by a considerable margin on the task of classifying windows as belonging to an object class or not. On object detection, this better window classifier leads to an improvement of about 3% mAP.

[1] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

[2] Y. Aytar and A. Zisserman. Enhancing exemplar svms using part level transfer regularization. In *BMVC*, 2012.

[3] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan. Subcategory-aware object classification. In *CVPR*, 2013.

[4] I. Endres, K. Shih, J. Jiaa, and D. Hoiem. Learning collections of part models for object recognition. In *CVPR*, 2013.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[6] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *CVPR*, 2013.

[7] M. Juneja, A. Vedaldi, CV Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[8] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[9] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. In *SIGGRAPH Asia Conference*, 2011.

[10] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[11] S. Song and J. Xiao. Sliding shapes for 3d object detection in depth images. In *ECCV*, 2014.

[12] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[13] A. Vezhnevets and V. Ferrari. Associative embeddings for large-scale knowledge transfer with self-assessment. In *CVPR*, 2014.