MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking - Supplementary

Thibaut Durand, Nicolas Thome, Matthieu Cord
Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu, 75005 Paris
{thibaut.durand, nicolas.thome, matthieu.cord}@lip6.fr

A. MANTRA Model

A.1. Learning Scheme

We show here that the loss function $\ell(\mathbf{x}_i, \mathbf{y}_i)$ (Eq. (4) of the submitted paper) is an upper bound of $\Delta(\hat{\mathbf{y}}, \mathbf{y}_i)$, where $\mathbf{x}_i$ is the input, $\mathbf{y}_i$ is the ground truth, and $\hat{\mathbf{y}}$ the predicted output.

Proof:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}) \tag{1}$$

$$\Delta(\hat{\mathbf{y}}, \mathbf{y}_i) \leq \Delta(\hat{\mathbf{y}}, \mathbf{y}_i) + \underbrace{D_{\mathbf{w}}(x_i, \hat{\mathbf{y}}) - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)}_{\geq 0} \tag{2}$$

$$\leq \max_{\mathbf{y} \in \mathcal{Y}} \big[\Delta(\mathbf{y}, \mathbf{y}_i) + D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}) - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)\big] \tag{3}$$

$$\leq \ell(\mathbf{x}_i, \mathbf{y}_i) \tag{4}$$

This proves that $\ell(\mathbf{x}_i, \mathbf{y}_i)$ is an upper bound of $\Delta(\hat{\mathbf{y}}, \mathbf{y}_i)$.

A.2. Optimization

A.2.1 1-Slack Dual Formulation

First we write the Lagrangian of primal formulation (Eq. (6) of the submitted paper)

$$\mathcal{L}(\mathbf{w}, \xi, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + C\xi - \alpha'\xi \tag{5}$$

$$- \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \left(\xi - \frac{1}{N} \sum_{i=1}^{N} \Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i) + D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \right)$$

where $\alpha' \geq 0$ and $\forall \bar{\mathbf{y}} = (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N) \in \mathcal{Y}^N$, $\alpha_{\bar{\mathbf{y}}} \geq 0$. Then, we differentiate the constraints with respect to the primal variables:

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \xi, \alpha) = \mathbf{w} \tag{6}$$

$$+ \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)$$

$$= 0 \tag{7}$$

The equation of $\mathbf{w}$ with dual variables is:

$$\mathbf{w} = - \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \tag{8}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, \xi, \alpha)}{\partial \xi} = C - \alpha' - \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} = 0 \tag{9}$$

This differentiation gives a condition on the sum of dual variables:

$$0 \leq \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \leq C \tag{10}$$

Dual formulation Applying the Eq. (8,9), in the Lagrangian (Eq. (6)), the dual formulation of the optimization problem (Eq. (6) of the submitted paper) is

$$\mathcal{D}(\alpha) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{11}$$

$$+ \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i) + D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)$$

$$= \frac{1}{2} \left\langle \mathbf{w}, - \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \right\rangle \tag{12}$$

$$+ \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i) + D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)$$

$$= -\frac{1}{2} \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \langle \mathbf{w}, \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \rangle \tag{13}$$

$$+ \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} (\Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i) + \langle \mathbf{w}, \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \rangle) \tag{14}$$

$$= \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i) \tag{15}$$

$$+ \frac{1}{2} \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \langle \mathbf{w}, \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \rangle$$

$$= \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i) \tag{16}$$

$$- \frac{1}{2} \left\langle \mathbf{w}, - \sum_{\bar{\mathbf{y}} \in \mathcal{Y}^N} \alpha_{\bar{\mathbf{y}}} \frac{1}{N} \sum_{i=1}^{N} \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \right\rangle$$

The Eq. (16) can be rewritten in the standard formulation

$$\alpha^T c - \frac{1}{2} \alpha^T H \alpha \tag{17}$$

where $\forall \bar{\mathbf{y}}, \bar{\mathbf{y}}' \in \mathcal{Y}^N \quad H_{\bar{\mathbf{y}}\bar{\mathbf{y}}'} = \langle g^{(\bar{\mathbf{y}})}, g^{(\bar{\mathbf{y}}')} \rangle$ with

$$g^{(\bar{\mathbf{y}})} = -\frac{1}{N} \sum_{i=1}^{N} \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \tag{18}$$

and $c_{\bar{\mathbf{y}}} = \frac{1}{N} \sum_{i=1}^{N} \Delta(\mathbf{y}_i, \hat{\mathbf{y}}_i)$

We solve the QP problem, with an interior-point optimizer, as in [1].

A.2.2 Detailed MANTRA Algorithm

Gradient computation We give the computation of the gradient of $\nabla_w \ell_w$ required in Algrithm 1 of the submitted paper (Line 13):

$$\nabla_{\mathbf{w}} \ell_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) = \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}) - \nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \tag{19}$$

where

$$\nabla_{\mathbf{w}} D_{\mathbf{w}}(\mathbf{x}_i, \hat{\mathbf{y}}) = \Psi(\mathbf{x}_i, \hat{\mathbf{y}}, \mathbf{h}^{+}_{i,\hat{\mathbf{y}}}) + \Psi(\mathbf{x}_i, \hat{\mathbf{y}}, \mathbf{h}^{-}_{i,\hat{\mathbf{y}}}) \tag{20}$$

B. Ranking Instantiation

B.1. Proof of Lemma 1

First, we remind the joint feature map used in section 4.2 of the submitted paper:

$$\Psi(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \mathbf{y}_{ij} (\Phi(x_i, h_{i,j}) - \Phi(x_j, h_{j,i})) \tag{21}$$

In this section, we prove that $D_{\mathbf{w}}(\mathbf{x}, \mathbf{y})$ can be re-written as a supervised feature map, where the score of each example x_i is $\langle \mathbf{w}, \Phi^{+}_{-}(x_i) \rangle$.

Given an input $\mathbf{x}$, and an output $\mathbf{y}$ and a weight vector $\mathbf{w}$, we have:

$$D_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{h} \in \mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}, \mathbf{h}) \rangle + \min_{\mathbf{h}' \in \mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}, \mathbf{h}') \rangle$$

$$= \max_{\mathbf{h} \in \mathcal{H}} \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h_{i,j}) \rangle - \langle \mathbf{w}, \Phi(x_j, h_{j,i}) \rangle)$$

$$+ \min_{\mathbf{h}' \in \mathcal{H}} \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h'_{i,j}) \rangle - \langle \mathbf{w}, \Phi(x_j, h'_{j,i}) \rangle) \tag{22}$$

$$= \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \max_{(h_i, h_j) \in \mathcal{H}_i \times \mathcal{H}_j} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h_i) \rangle - \langle \mathbf{w}, \Phi(x_j, h_j) \rangle)$$

$$+ \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \min_{(h'_i, h'_j) \in \mathcal{H}_i \times \mathcal{H}_j} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h'_i) \rangle - \langle \mathbf{w}, \Phi(x_j, h'_j) \rangle) \tag{23}$$

With the definition of the latent variable and the joint feature, the maximization (resp. minimization) over the latent variables can be decomposed for each term of the sum. So maximizing (resp. minimizing) the sum is equivalent to maximize (resp. minimize) each term of the sum independently, because the latent variable $\mathbf{h}$ can be decomposed for each term of the sum and each couple of latent variables $(h_{i,j}, h_{j,i})$ is independent.

Now, the 2 sums are grouped in a single sum:

$$\frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \max_{(h_i, h_j) \in \mathcal{H}_i \times \mathcal{H}_j} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h_i) \rangle - \langle \mathbf{w}, \Phi(x_j, h_j) \rangle)$$

$$+ \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \min_{(h'_i, h'_j) \in \mathcal{H}_i \times \mathcal{H}_j} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h'_i) \rangle - \langle \mathbf{w}, \Phi(x_j, h'_j) \rangle) \tag{24}$$

$$= \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \Big(\max_{(h_i, h_j) \in \mathcal{H}_i \times \mathcal{H}_j} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h_i) \rangle - \langle \mathbf{w}, \Phi(x_j, h_j) \rangle)$$

$$+ \min_{(h'_i, h'_j) \in \mathcal{H}_i \times \mathcal{H}_j} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h'_i) \rangle - \langle \mathbf{w}, \Phi(x_j, h'_j) \rangle) \Big) \tag{25}$$

We define $A(\mathbf{x}, \mathbf{y}) =$

$$\frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{x_i \in \mathcal{P}} \sum_{x_j \in \mathcal{N}} \Big(\max_{(h_i, h_j) \in \mathcal{H}_i \times \mathcal{H}_j} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h_i) \rangle - \langle \mathbf{w}, \Phi(x_j, h_j) \rangle)$$

$$+ \min_{(h'_i, h'_j) \in \mathcal{H}_i \times \mathcal{H}_j} \mathbf{y}_{ij} (\langle \mathbf{w}, \Phi(x_i, h'_i) \rangle - \langle \mathbf{w}, \Phi(x_j, h'_j) \rangle) \Big) \tag{26}$$

By construction, we have the equality $A(\mathbf{x}, \mathbf{y}) = D_{\mathbf{w}}(\mathbf{x}, \mathbf{y})$. Now, we show that the latent variables can be fixed independently to the ranking matrix $\mathbf{y}$. For a couple of examples (x_i, x_j), with $x_i \in \mathcal{P}, x_j \in \mathcal{N}$, we analyze the value of the latent variables h_i, h_j, h'_i, h'_j with respect to $\mathbf{y}_{ij}$. There is only 2 cases to analyze: $\mathbf{y}_{ij} = 1$ and $\mathbf{y}_{ij} = -1$.

If $\mathbf{y}_{ij} = 1$

$$\max_{(h_i,h_j)\in\mathcal{H}_i\times\mathcal{H}_j}(\langle\mathbf{w},\Phi(x_i,h_i)\rangle-\langle\mathbf{w},\Phi(x_j,h_j)\rangle) \quad (27)$$
$$+\min_{(h'_i,h'_j)\in\mathcal{H}_i\times\mathcal{H}_j}(\langle\mathbf{w},\Phi(x_i,h'_i)\rangle-\langle\mathbf{w},\Phi(x_j,h'_j)\rangle)$$
$$=\langle\mathbf{w},\Phi(x_i,h_i^+)\rangle-\langle\mathbf{w},\Phi(x_j,h_j^-)\rangle \quad (28)$$
$$+\langle\mathbf{w},\Phi(x_i,h_i^-)\rangle-\langle\mathbf{w},\Phi(x_j,h_j^+)\rangle$$
$$=\langle\mathbf{w},\Phi(x_i,h_i^+)+\Phi(x_i,h_i^-)\rangle-\langle\mathbf{w},\Phi(x_j,h_j^+)+\Phi(x_j,h_j^-)\rangle \quad (29)$$
$$=\langle\mathbf{w},\Phi_-^+(x_i)\rangle-\langle\mathbf{w},\Phi_-^+(x_j)\rangle \quad (30)$$
$$\text{where}\quad \Phi_-^+(x_i)=\Phi(x_i,h_i^+)+\Phi(x_i,h_i^-) \quad (31)$$
$$h_i^+=\arg\max_{h\in\mathcal{H}_i}\langle\mathbf{w},\Phi(x_i,h)\rangle \quad (32)$$
$$h_i^-=\arg\min_{h\in\mathcal{H}_i}\langle\mathbf{w},\Phi(x_i,h)\rangle \quad (33)$$

If $\mathbf{y}_{ij} = -1$

$$\max_{(h_i,h_j)\in\mathcal{H}_i\times\mathcal{H}_j}-(\langle\mathbf{w},\Phi(x_i,h_i)\rangle-\langle\mathbf{w},\Phi(x_j,h_j)\rangle) \quad (34)$$
$$+\min_{(h'_i,h'_j)\in\mathcal{H}_i\times\mathcal{H}_j}-(\langle\mathbf{w},\Phi(x_i,h'_i)\rangle-\langle\mathbf{w},\Phi(x_j,h'_j)\rangle)$$
$$=\max_{(h_i,h_j)\in\mathcal{H}_i\times\mathcal{H}_j}(\langle\mathbf{w},\Phi(x_j,h_j)\rangle-\langle\mathbf{w},\Phi(x_i,h_i)\rangle) \quad (35)$$
$$+\min_{(h'_i,h'_j)\in\mathcal{H}_i\times\mathcal{H}_j}(\langle\mathbf{w},\Phi(x_j,h'_j)\rangle-\langle\mathbf{w},\Phi(x_i,h'_i)\rangle)$$
$$=\langle\mathbf{w},\Phi(x_j,h_j^+)\rangle-\langle\mathbf{w},\Phi(x_i,h_i^-)\rangle \quad (36)$$
$$+\langle\mathbf{w},\Phi(x_j,h_j^-)\rangle-\langle\mathbf{w},\Phi(x_i,h_i^+)\rangle$$
$$=-(\langle\mathbf{w},\Phi_-^+(x_i)\rangle-\langle\mathbf{w},\Phi_-^+(x_j)\rangle) \quad (37)$$

We notice that the predicted latent variables are the same in the two cases. So the latent variables can be fixed independently to the value of $\mathbf{y}_{ij}$.

$$\max_{(h_i,h_j)\in\mathcal{H}_i\times\mathcal{H}_j}\mathbf{y}_{ij}(\langle\mathbf{w},\Phi(x_i,h_i)\rangle-\langle\mathbf{w},\Phi(x_j,h_j)\rangle) \quad (38)$$
$$+\min_{(h'_i,h'_j)\in\mathcal{H}_i\times\mathcal{H}_j}\mathbf{y}_{ij}(\langle\mathbf{w},\Phi(x_i,h'_i)\rangle-\langle\mathbf{w},\Phi(x_j,h'_j)\rangle)$$
$$=\mathbf{y}_{ij}(\langle\mathbf{w},\Phi_-^+(x_i)\rangle-\langle\mathbf{w},\Phi_-^+(x_j)\rangle)$$

When the latent variables are fixed, each example x_i can be represented by $\Phi_-^+(x_i)$, and $A(\mathbf{x},\mathbf{y})$ can be written as follow:

$$A(\mathbf{x}\mathbf{y})=\frac{1}{|\mathcal{P}||\mathcal{N}|}\sum_{x_i\in\mathcal{P}}\sum_{x_j\in\mathcal{N}}\mathbf{y}_{ij}(\langle\mathbf{w},\Phi_-^+(x_i)\rangle-\langle\mathbf{w},\Phi_-^+(x_j)\rangle) \quad (39)$$

So $D_{\mathbf{w}}(\mathbf{x},\mathbf{y})$ can be written as a supervised feature map, where the latent are fixed independently to the ranking matrix $\mathbf{y}$, and each example x_i is represented by $\Phi_-^+(x_i)$.

B.2. Complexity Analysis

In this section, we analyze the complexity of inference and loss-augmented inference for ranking instantiation (section 4.2 of the submitted paper). We define $\bar{h}$ as the average number of regions per images.

Inference The inference complexity with MANTRA is $O(N\bar{h}d + N\log N)$, where the first term is the complexity to infer exhaustively the latent variables and the second term is the complexity of the sort.

Loss-augmented inference To solve loss-augmented inference with AP loss, we use the algorithm proposed by [2]. With this algorithm, the complexity of loss-augmented inference for MANTRA is $O(N\bar{h}d + N\log N + |\mathcal{P}||\mathcal{N}|)$, where the third term is the complexity to rank negative examples.

C. Experiments

C.1. Multi-class experiments

Datasets First, we detail here the evaluation protocol for the 4 databases (15-Scene, PPMI, MIT 67 Indoor Scenes, UIUC-Sports) used in Section 5.1 of the submitted paper. For 15-Scene (resp. UIUC-Sports), with 4485 (resp. 1574) images in total, 100 (resp. 70) examples for each class are randomly sampled for training, the remainder (resp. 60) being used for testing. As commonly done, we randomly sample the train/test folds 5 times and report the mean accuracy over the 5 runs. For MIT67 (resp. PPMI), we used the pre-defined split of the data, with 80 (resp. 100) training images and 20 (resp. 100) testing images for each category. The number of categories for UIUC Sports (resp. 15 Scene, PPMI, MIT67) is 8 (resp. 15, 24, 67). The evaluation metric in all datasets is the multi-class accuracy.

Results Table 1 (resp. Table 2) gives the values of Figure 2 (resp. 3) of the submitted paper. Figure 1 shows the training time results for 15 Scene and PPMI. For 15 Scene and MIT67, the results for scale 100% slightly differ from [3], mainly because [3] uses a non-standard evaluation protocol. [4] proposes a method to learn discriminative and shareable features (DSFL). Results are from 10 to 20 pt below our results, and also below the deep features baseline. Authors only show good results when they combine with deep features, but always lower than our results. As they claim complementarity between DSFL and deep features, we tried to combine DSFL with MANTRA. We have re-implemented the DSFL method, but we could not replicate their combination results. The authors did not answer our e-mails and source code has still not been released. So,

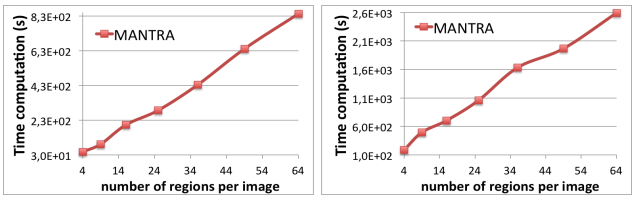
Scale	100	90	80	70	60	50	40	30
UIUC-Sports	94.4 ± 0.7	95.7 ± 0.7	96.4 ± 0.7	96.2 ± 0.6	95.8 ± 0.4	95.6 ± 0.6	94.5 ± 0.5	93.2 ± 1
15 Scene	90.7 ± 0.5	91.7 ± 0.2	92.2 ± 0.3	91.2 ± 0.2	90.7 ± 0.3	88.9 ± 0.3	85.4 ± 1	80.7 ± 0.7
PPMI	54.5	56.1	56.9	58.6	58.9	59.2	54.7	51.0
MIT67	69.9	71.8	72.6	72.1	71	66.4	63	56.4

Table 1. Performances of MANTRA for the different scales.

we did not consider DSFL in our submission as an option for comparison and further combination.

Scale	90	80	70	60	50	40	30
Regions	4	9	16	25	36	49	64
UIUC	5	11	18	27	37	47	61
15 Scene	48	92	203	287	434	641	843
PPMI	187	495	705	1058	1632	1697	2593
MIT67	2749	6443	11200	16356	24029	31705	41805

Table 2. Time computation (in seconds). The second row (regions) is the number of regions per image



15 Scene PPMI

Figure 1. MANTRA training time (seconds) *vs* number of region per image (seconds - values are reported in Table 2)

Visual results In Figure 2 (resp. Figure 3, Figure 4), we show visual results on UIUC-Sports (resp. MIT67 and 15 Scene) dataset. We show prediction maps and $(\mathbf{h}^+, \mathbf{h}^-)$ regions, for different classifiers. For instance, when classifying a snowboard image (Figure 2, last row), the model learns that the detection of snow supports the absence of other categories, e.g. polo, bocce, or croquet. For the correct class, the incorporation of $\mathbf{h}^-$ prevents from having large negative values for any (random) window, thus $\mathbf{h}^-$ can be regarded as a regularizer on the latent space exploiting contextual information. The fourth row of Figure 2 shows the prediction of *sailing* and *rowing* categories for a *sailing* image. For each classifier, $\mathbf{h}^+$ corresponds to discriminative parts, *i.e.* boat with sail and water. The $\mathbf{h}^-$ region for the *rowing* classifier focuses on the sail of the boat with a very low score. It suggests that if a sail is found, the image is very unlikely to belong to the class *rowing*. Another example is the *greenhouse* image of MIT67 (second row of Figure 3), where both *greenhouse* and *florist* classifiers have high scores and focus on plants. For the *greenhouse* classifier, $\mathbf{h}^-_{greenhouse}$ has a quite high score $(+0.8)$, so all regions are discriminative for it. This is in stark contrast to the *florist* classifier, for which $\mathbf{h}^-_{florist}$ has a very low score (-0.9): it thus finds clear evidence of the absence of the *florist* category. For classifiers of un-correlated categories like *laboratorywet*, all regions have very low scores (< -0.9).

In Figures 2, 3, 4, we can point out other examples of fine-grained classification problems, where wrong classifiers can have large scores on local regions, which are, however, compensated by very strong evidence of the absence of the class: *croquet vs bocce* in Figure 2 (UIUC-Sports), *closet vs clothing store* and *book store vs library* in Figure 3 (MIT67), *street vs inside city* or *highway* and *tall building vs inside city* in Figure 4 (15 Scene).

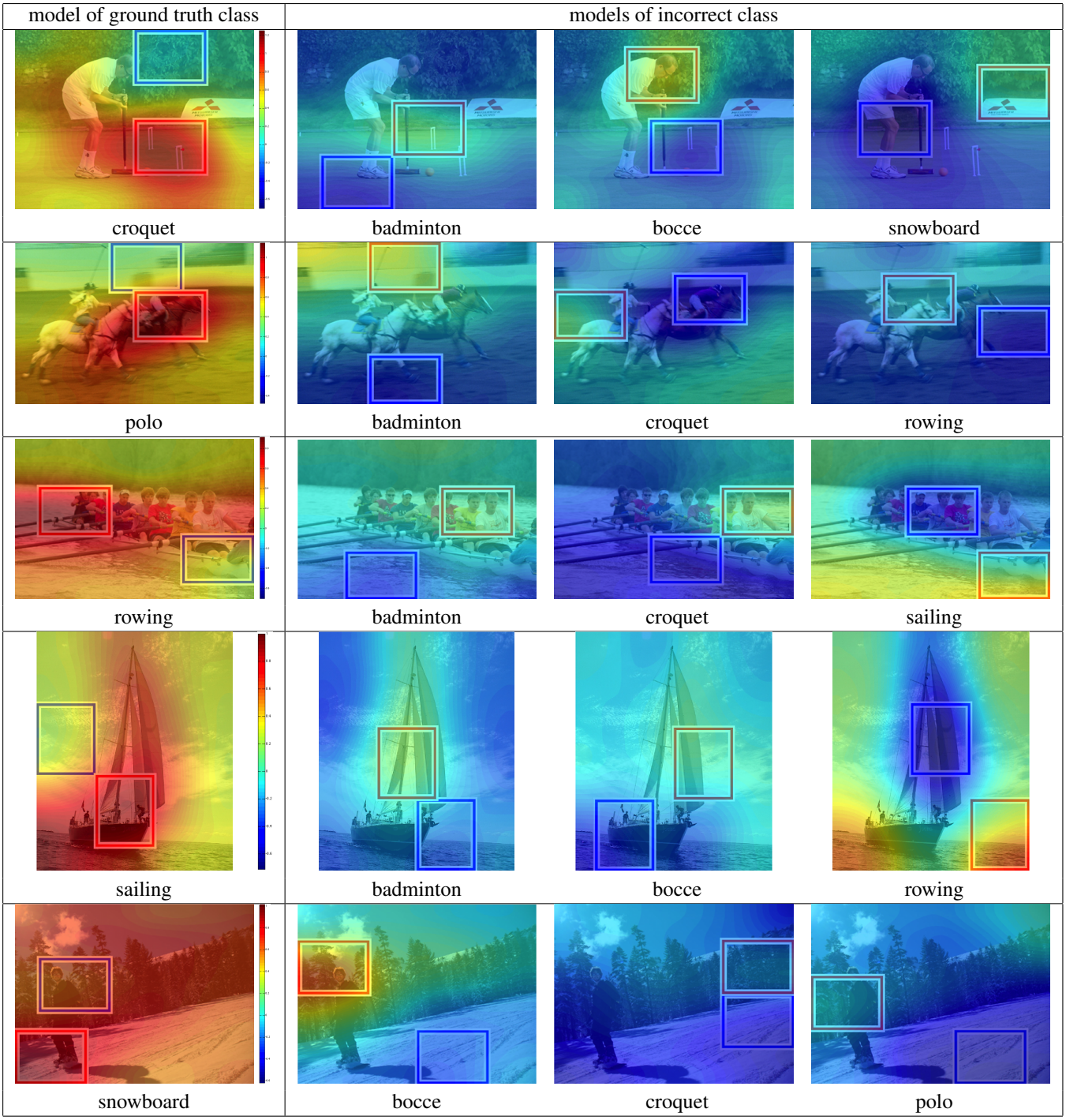


Figure 2. Example of response map for UIUC-Sports images and for model of the correct class (left column) and models of incorrect class. For each model, the red (resp. blue) bounding box show the region with the maximum (resp. minimum) score.

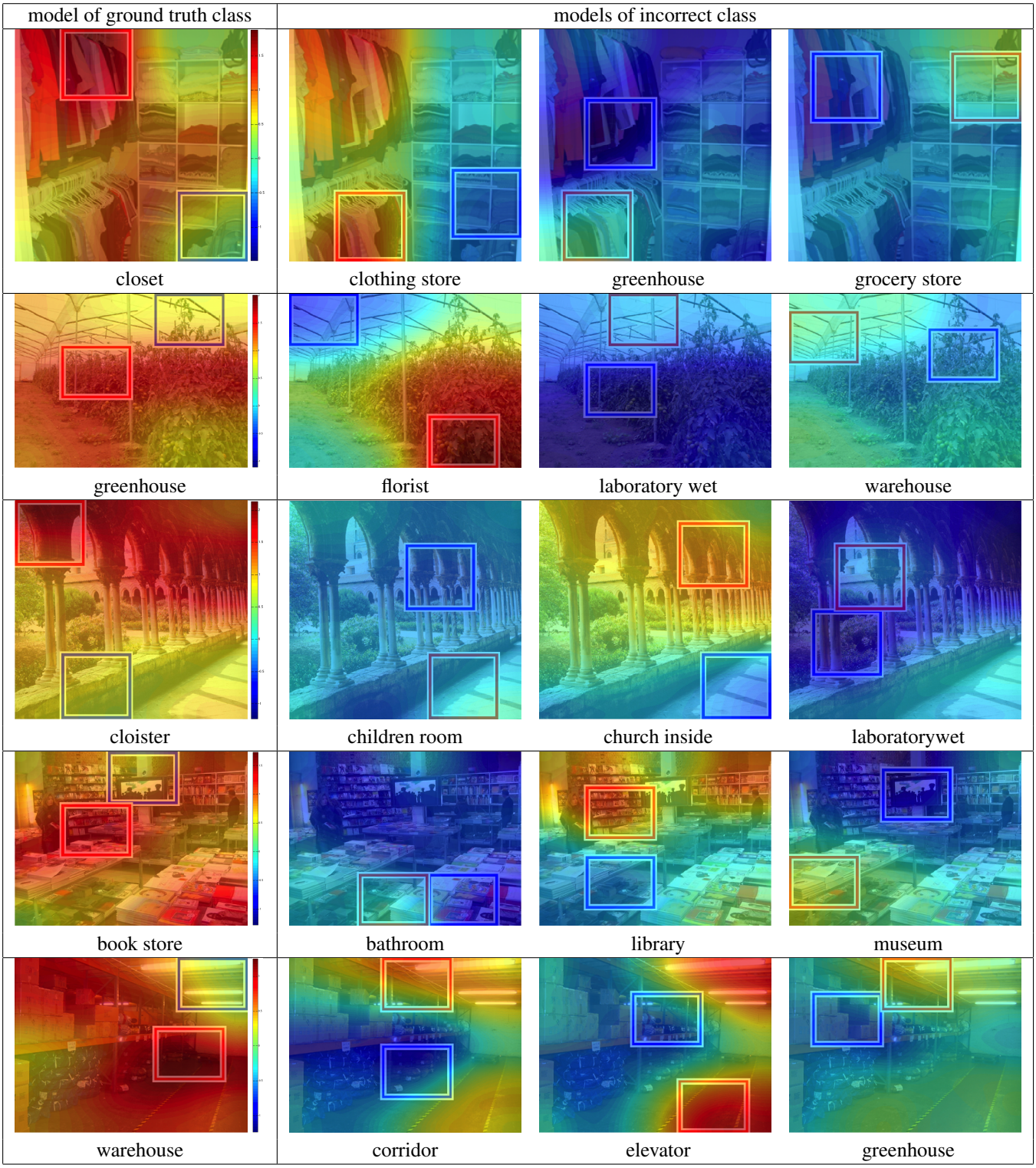


Figure 3. Example of response map for MIT67 images and for model of the correct class (left column) and models of incorrect class. For each model, the red (resp. blue) bounding box show the region with the maximum (resp. minimum) score.

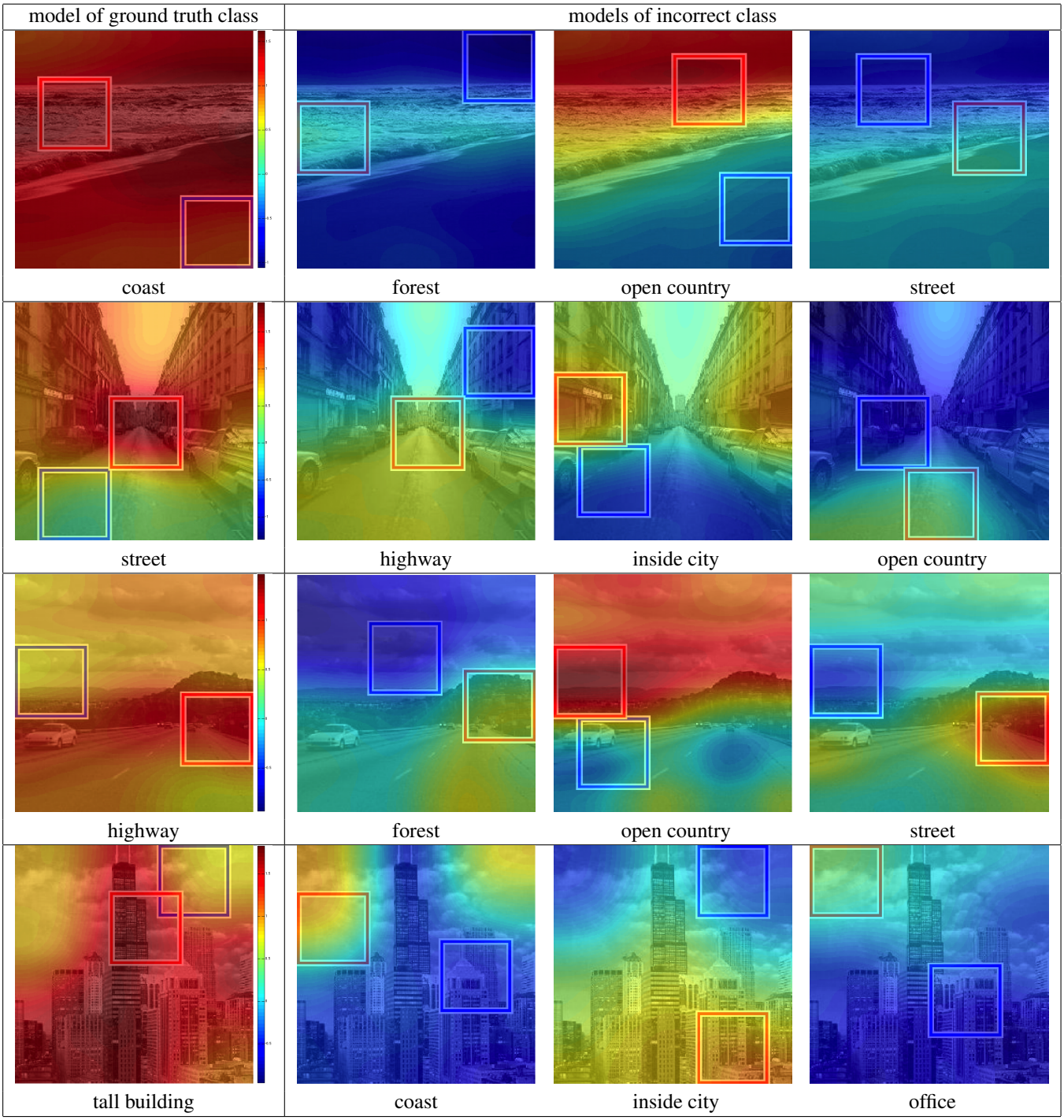


Figure 4. Example of response map for 15 Scene images and for model of the correct class (left column) and models of incorrect class. For each model, the red (resp. blue) bounding box show the region with the maximum (resp. minimum) score.

C.2. Ranking Results on VOC 2011 Action

In this section, we details the ranking and detection performances. In Table 3, ranking performances are reported for the 5 splits. The paired T-test between LAPSVM and MANTRA-AP ranking performances is 8.98, so the difference is significant with a risk of 0.1% (critical value is 8.6101). We also measure the detection performances on the testing set, by computing the average overlap (intersection over union) between the predicted region and the ground truth bounding box of the person. Table 4 provides detection performances. The paired T-test between LAPSVM and MANTRA-AP detection performances is 9.48, so the difference is significant with a risk of 0.1%.

Split	1	2	3	4	5
LSSVM-Acc	28,4	28.6	29.5	25.9	28.8
MANTRA-Acc	37.0	36.4	32.6	33.7	36.2
LAPSVM	37.8	36.4	37.3	35.4	36.7
MANTRA-AP	44.1	41.6	40.5	41.6	43.1

Table 3. Ranking: Mean Average Precision on VOC 2011 Action Classification dataset for the 5 splits.

Split	1	2	3	4	5
LSSVM-Acc	12.5	12.9	12.6	12.8	12.7
MANTRA-Acc	19.2	18.7	19.3	19	18.4
LAPSVM	19.2	19.6	20.9	20.8	20.2
MANTRA-AP	26	25.7	24.8	28	28

Table 4. Detection: mean overlap on VOC 2011 Action Classification dataset for the 5 splits.

References

[1] C.-N. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, 2009. 2

[2] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *SIGIR*, 2007. 3

[3] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014. 3

[4] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang. Learning discriminative and shareable features for scene classification. In *ECCV*, 2014. 3