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

Thibaut DURAND, Nicolas THOME, Matthieu CORD
Sorbonne Universités, UPMC Univ Paris 06, LIP6, Paris, France

DECEMBER 11-18, 2015
ICCV15
International Conference on Computer Vision

Context

x

y=snowboarding

- Supervision: full annotations (*e.g.* BB) expensive
- Weakly Supervised Learning (WSL) framework
  - Option: using latent variables
  - Most popular framework: LSSVM [1]
  - Ranking optimization challenging [2]

Contributions

- MANTRA: new structured output latent model
  - 2 latent variables: max + min
- Efficient optimization
  - 2 instantiations: multi-class, AP ranking
- Experimental validation on 6 datasets

MANTRA Model

- **Scoring function:**

$$D_{\mathbf{w}}(\mathbf{x},\mathbf{y}) = \underbrace{\langle \mathbf{w}, \Psi(\mathbf{x},\mathbf{y},\mathbf{h}^+_{\mathbf{y}})\rangle}_{s(\mathbf{h}^+_{\mathbf{y}})} + \underbrace{\langle \mathbf{w}, \Psi(\mathbf{x},\mathbf{y},\mathbf{h}^-_{\mathbf{y}})\rangle}_{s(\mathbf{h}^-_{\mathbf{y}})}$$

- **Prediction function:**

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} D_{\mathbf{w}}(\mathbf{x},\mathbf{y})$$

- **Notations:**
  - **max** scoring latent value

$$\mathbf{h}^+_{\mathbf{y}} = \arg\max_{\mathbf{h}} \langle \mathbf{w}, \Psi(\mathbf{x},\mathbf{y},\mathbf{h})\rangle$$

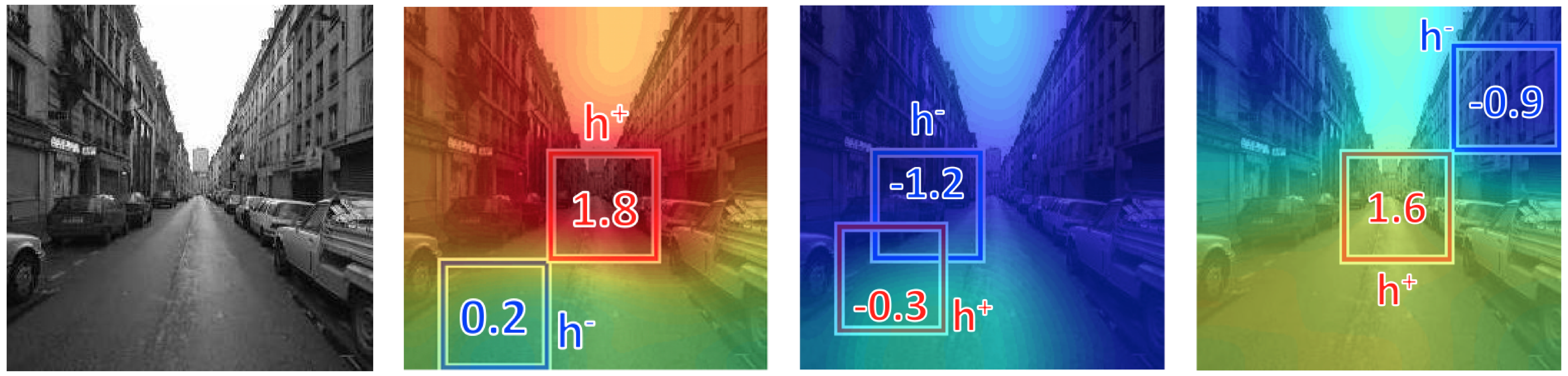
  - **min** scoring latent value

$$\mathbf{h}^-_{\mathbf{y}} = \arg\min_{\mathbf{h}} \langle \mathbf{w}, \Psi(\mathbf{x},\mathbf{y},\mathbf{h})\rangle$$

  - $\mathbf{x}$: input (image)
  - $\mathbf{y}$: output (multi-class label, ranking matrix)
  - $\mathbf{h}$: latent (bounding box)
  - $\Psi(\mathbf{x},\mathbf{y},\mathbf{h}) \in \mathbb{R}^d$: joint feature map
  - $\mathbf{w} \in \mathbb{R}^d$: model parameters

Intuition

- $s(\mathbf{h}^+_{\mathbf{y}})$: witnesses the <u>**presence**</u> of the class
- $s(\mathbf{h}^-_{\mathbf{y}})$: witnesses the <u>**absence**</u> of the class
- $\mathbf{h}^-_{\mathbf{y}}$: contextual information complementary to $\mathbf{h}^+_{\mathbf{y}}$ (latent space regularizer)



original image	$D_{\mathbf{w}}(\mathbf{x},\textbf{street}) = 2$	$D_{\mathbf{w}}(\mathbf{x},\textbf{coast}) = -1.5$	$D_{\mathbf{w}}(\mathbf{x},\textbf{highway}) = 0.7$
correct class: **street**	$s(\mathbf{h}^+_{\textbf{street}}) = 1.8$: high	$s(\mathbf{h}^+_{\textbf{coast}}) = -0.3$: low	$s(\mathbf{h}^+_{\textbf{highway}}) = 1.6$: high
	$s(\mathbf{h}^-_{\textbf{street}}) = 0.2$: medium	$s(\mathbf{h}^-_{\textbf{coast}}) = -1.2$: low	$s(\mathbf{h}^-_{\textbf{highway}}) = -0.9$: low

Prediction: $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} D_{\mathbf{w}}(\mathbf{x},\mathbf{y}) \Rightarrow$ **street**

Experiments

Multi-class

UIUC 15Scene PPMI MIT67

- **Features:** Multi-scale deep features (Caffe)
- **Comparison to state-of-the-art models**

	UIUC	15Sc	PPMI	MIT67
num. classes	8	15	24	67
ImageNet	94	88	54.5	58.5
Places [4]	94.1	90.2	38.6	68.2
MOP-CNN [5]	-	-	-	68.9
MANTRA	**97.3**	**93.4**	**66.2**	**76.6**

- **Comparison to LSSVM**

Mono-scale results for the smallest scale

	UIUC	15-Scene	PPMI	MIT67
LSSVM [1]	73.3	65	13.3	26.6
MANTRA	**93.2**	**80.7**	**51.0**	**56.4**

Ranking

VOC 2007 VOC 2011 Action

- **VOC 2007 Results**

	[3]	MANTRA-Acc	MANTRA-AP
MAP(%)	82.4	82.6	**85.8**

- **VOC 2011 Action Results**

Method	Ranking AP (%)	Detect. ov. (%)
LSSVM [1]	29.5 ± 1.3	12.7 ± 0.3
MANTRA-Acc	35.2 ± 1.2	18.9 ± 0.9
LAPSVM [2]	36.7 ± 0.8	20.1 ± 0.7
MANTRA-AP	**42.2 ± 1.3**	**26.5 ± 1.4**

[1] C.-N. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, 2009.
[2] Behl *et al.* Optimizing average precision using weakly supervised data. *CVPR*, 2014.
[3] Chatfield *et al.* Return of the Devil in the Details: Delving Deep into Convolutional Nets *BMVC*, 2014.
[4] Zhou *et al.* *NIPS*, 2014.
[5] Gong *et al.* *ECCV*, 2014.

Learning

- N training pairs $(\mathbf{x}_i, \mathbf{y}_i)$
- $\Delta(\mathbf{y}_i, \mathbf{y})$: user-defined loss function
- **Constraints: during training:**

$$\forall \mathbf{y} \neq \mathbf{y}_i, \quad D_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i) \geq \Delta(\mathbf{y}_i,\mathbf{y}) + D_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y})$$

- **Primal objective:**

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} \max_{\mathbf{y}}[\Delta(\mathbf{y}_i,\mathbf{y}) + D_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y})] - D_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)$$

- **Optimization:** non-convex cutting plane

MANTRA Instantiation

- Define feature map Ψ and loss function Δ.
- Solve inference and loss-augmented inference (LAI) (during training):

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} [\Delta(\mathbf{y}_i,\mathbf{y}) + D_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y})]$$

	Multi-class	**Ranking AP**
$\mathbf{x}$	image	set of images
$\mathbf{y}$	multi-class label	ranking matrix
$\mathbf{h}$	region	regions
$\Psi(\mathbf{x},\mathbf{y},\mathbf{h})$	joint multi-class feature map	joint latent ranking feature map
$\Delta(\mathbf{y}_i,\mathbf{y})$	0/1 loss	AP loss
LAI	exhaustive	exact and efficient

- **MANTRA ranking**: **exact** and **efficient solutions** for inference and LAI (proof in the paper)
  - decoupling the optimization over $\mathbf{y}$ and $\mathbf{h}$

Conclusion

- max + min scoring function ≫ max
- AP optimization: **significant improvements**
- **State-of-the-art results** on 5 datasets

rowing croquet sailing

- **Code available project page: give it a try!**

`http://webia.lip6.fr/~durandt/project/mantra.html`