Weakly Supervised Learning of Deep Structured Models

Nicolas Thome

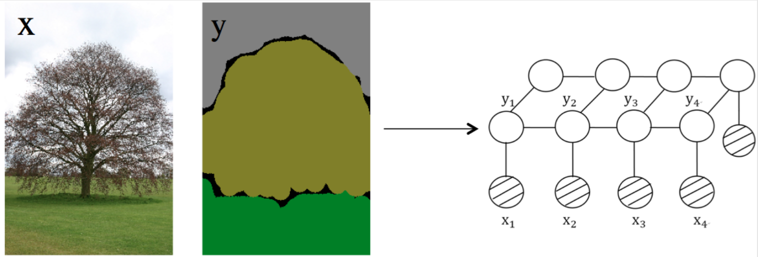
Université Pierre et Marie Curie (UPMC)
Laboratoire d'Informatique de Paris 6 (LIP6)

Outline

Structured prediction

Structured inputs and outputs

- $\mathcal{X}$ is the input space : arbitrary (non vectorial, *etc*)
- $\mathcal{Y}$ is the structured output space: discrete, with variables strongly correlated ⇒ probabilistic graphical models (chain, tree, general graph)
- Ex: semantic image segmentation ⇒ classify each pixel into semantic categories
  - Output space $\mathcal{Y} = \{1, ..., k\}^D$ with correlated variables



Structured prediction

Structural SVM (SSVM) [TJHA05]

- Relationship between input $\mathbf{x} \in \mathcal{X}$ and output $\mathbf{y} \in \mathcal{Y}$
  $\Rightarrow$ joint feature map $\Psi(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$
- Scoring function linear in Ψ: $f_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) = \langle w, \Psi(\mathbf{x}, \mathbf{y}) \rangle = s(\mathbf{y})$
  - Kernel extension possible
  - $\Psi(\mathbf{x}, \mathbf{y})$ possibly deep, WELDON or [CSYU15]
- Prediction or **inference**:
  $$\hat{y}(x, w) = \arg\max_{y \in \mathcal{Y}} \langle w, \Psi(x, y) \rangle = \arg\max_{y \in \mathcal{Y}} s(\mathbf{y})$$
- Output space $\mathcal{Y}$ generally huge $\Rightarrow$ exhaustive maximization not tractable
  - Exploit structure (exact solutions for chain, trees), specific scoring functions (sub-modular), *etc*
    - Inference in graphical models: extremely rich literature

Structured prediction

Structural SVM: training

- Training: a set of N labeled trained pairs $(\mathbf{x}_i, \mathbf{y}_i)$
- Structured loss $\Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i)$, $\hat{\mathbf{y}}_i(\mathbf{x}_i, \mathbf{w}) = \arg\max_{y \in \mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle$

  $\Rightarrow$ *Prior* (expert) knowledge on the dissimilarity between two outputs
- Dependence of Δ wrt $\mathbf{w}$ complex (non-convex, non-smooth)
- **Margin rescaling:** convex upper bound $\Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i) \leq \ell(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})$

  $\ell(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}) = \max_{\mathbf{y} \in \mathcal{Y}} \left[\Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle\right] - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle$
- $\max_{\mathbf{y} \in \mathcal{Y}} \left[\Delta(\mathbf{y}_i, \mathbf{y}) + s(\mathbf{y})\right]$ "Loss Augmented Inference" (LAI) $\Rightarrow$ exhaustive maximization not tractable
  - Generally harder than inference (depends on Δ)

Structured prediction

Structured Output Ranking

- Input $\mathcal{X}$ list of n examples: $\mathbf{x} = (o_1, ...o_n)$
- Output $\mathcal{Y}$ ranking of examples ($|\mathcal{Y}| \sim 2^{n^2/2}$): $\mathbf{y}$ matrix s.t.

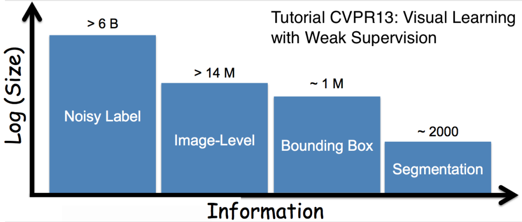
$$y_{ij} = \begin{cases} +1 & \text{if } o_i \prec_y o_j \text{ (} o_i \text{ is before } o_j \text{ in the sorted list)} \\ -1 & \text{if } o_i \succ_y o_j \text{ (} o_i \text{ is after } o_j \end{cases}$$

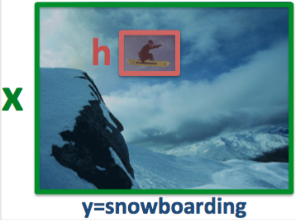
- Ranking feature map: $\Psi(\mathbf{x}, \mathbf{y}) = \sum_{i \in \oplus} \sum_{j \in \ominus} y_{ij} \left(\phi(o_i) - \phi(o_j) \right)$
- **Inference:** exact by sorting example wrt $\langle \mathbf{w}; \phi(o_i) \rangle$ [YFRJ07]
- **LAI** with Average Precision (AP) loss: $\Delta_{AP}(y_i, y) = 1 - AP(y)$
  - Δ_{AP}: no linear decomposition wrt examples $\neq$ AUC (ROC)
  - Optimal greedy algorithm in $O(nlog(n))$ [YFRJ07]
  - Speed-up in NIPS'14 [MJK14]

Structured prediction with latent variables

Weakly Supervised Learning (WSL)

- Full annotations expensive ⇒ training with weak supervision





- Incorporating latent variables $\mathbf{h} \in \mathcal{H}$

Variable	Notation	Space	Train	Test
Input	$\mathbf{x}$	$\mathcal{X}$	observed	observed
Output	$\mathbf{y}$	$\mathcal{Y}$	observed	unobserved
Latent	$\mathbf{h}$	$\mathcal{H}$	unobserved	unobserved

Structured prediction with latent variables

Latent Structural SVM [YJ09]

- Prediction function : $(\hat{\mathbf{y}}, \hat{\mathbf{h}}) = \underset{(\mathbf{y},\mathbf{h}) \in \mathcal{Y} \times \mathcal{H}}{\arg\max} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}, \mathbf{h}) \rangle = \underset{(\mathbf{y},\mathbf{h}) \in \mathcal{Y} \times \mathcal{H}}{\arg\max} s(\mathbf{y}, \mathbf{h})$
  - Joint inference in the $(\mathcal{Y} \times \mathcal{H})$ space
- Training: a set of N labeled trained pairs $(\mathbf{x}_i, \mathbf{y}_i)$
- Training objective: upper bound of $\Delta(\hat{\mathbf{y}}_i, \mathbf{y}_i)$:

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N} \sum_{i=1}^{N} \max_{(\mathbf{y},\mathbf{h}) \in \mathcal{Y} \times \mathcal{H}} \left[\Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h}) \rangle\right] - \max_{\mathbf{h} \in \mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}) \rangle$$

  - Difference of Convex Functions, solved with CCCP
  - LAI: $\max_{(\mathbf{y},\mathbf{h}) \in \mathcal{Y} \times \mathcal{H}} \left[\Delta(\mathbf{y}_i, \mathbf{y}) + s(\mathbf{y}, \mathbf{h})\right]$
    - Challenge exacerbated in the latent case, $(\mathcal{Y} \times \mathcal{H})$ space
    - No exact solution for structured AP ranking [BMJK15]
      $\Rightarrow$ Approximate solution in [BMJK15]

Outline

MANTRA: Minimum Maximum Latent Structural SVM

MANTRA model

- Pair of latent variables $(\mathbf{h}^+_{i,\mathbf{y}}, \mathbf{h}^-_{i,\mathbf{y}})$
  - **max** scoring latent value: $\mathbf{h}^+_{i,\mathbf{y}} = \underset{\mathbf{h}\in\mathcal{H}}{\arg\max}\ \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h})\rangle$
  - **min** scoring latent value: $\mathbf{h}^-_{i,\mathbf{y}} = \underset{\mathbf{h}\in\mathcal{H}}{\arg\min}\ \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h})\rangle$
- New scoring function:

$$\begin{aligned} D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}) &= \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h}^+_{i,\mathbf{y}})\rangle + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h}^-_{i,\mathbf{y}})\rangle \\ &= s(\mathbf{y}, \mathbf{h}^+_{\mathbf{y}}) + s(\mathbf{y}, \mathbf{h}^-_{\mathbf{y}}) \end{aligned} \tag{1}$$

- **MANTRA:** `max+min` *vs* `max` for **LSSVM** ⇒ **negative evidence**
- Prediction function ⇒ find the output with maximum score

$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in\mathcal{Y}}{\arg\max}\ D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}) \tag{2}$$

MANTRA: Model Training

Learning formulation

- Loss function: $\ell_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) = \max\limits_{\mathbf{y} \in \mathcal{Y}} \left[\Delta(\mathbf{y}_i, \mathbf{y}) + D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y})\right] - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)$
  - (Margin rescaling) upper bound of $\Delta(\mathbf{y}_i, \hat{\mathbf{y}})$, constraints:

$$\forall \mathbf{y} \neq \mathbf{y}_i, \quad \underbrace{D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)}_{\text{score for ground truth output}} \geq \underbrace{\Delta(\mathbf{y}_i, \mathbf{y})}_{\text{margin}} + \underbrace{D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y})}_{\text{score for other output}}$$

- Non-convex optimization problem

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N} \sum_{i=1}^{N} \ell_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \tag{3}$$

- Solver: non convex one slack cutting plane [DA12]
  - Fast convergence
  - Direct optimization $\neq$ CCCP for LSSVM
  - Still needs to solve LAI: $\max_y \left[\Delta(\mathbf{y}_i, \mathbf{y}) + D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y})\right]$

MANTRA: Model & Training Rationale

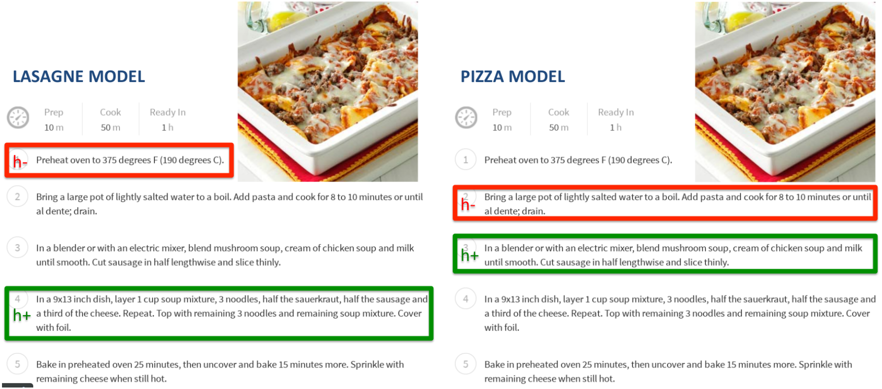
Intuition of the `max+min` prediction function

- $\mathbf{x}$ image, $\mathbf{h}$ image region, $\mathbf{y}$ image class
- $\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h}) \rangle = s(\mathbf{y}, \mathbf{h})$: region $\mathbf{h}$ score for class $\mathbf{y}$: heatmap
- $s(\mathbf{y}) = s(\mathbf{y}, \mathbf{h}^+_{\mathbf{y}}) + s(\mathbf{y}, \mathbf{h}^-_{\mathbf{y}})$
  - $\mathbf{h}^+_{\mathbf{y}}$: **presence** of class $\mathbf{y}$ $\Rightarrow$ large for $\mathbf{y}_i$
  - $\mathbf{h}^-_{\mathbf{y}}$: **localized evidence of the absence of class y**
    - Not too low for $\mathbf{y}_i$ $\Rightarrow$ latent space regularization
    - Low for $\mathbf{y} \neq \mathbf{y}_i$ $\Rightarrow$ tracking negative evidence [PVZF15]

street image $\mathbf{x}$ $\quad D_{\mathbf{w}}(\mathbf{x}, \textbf{street}) = 2$ $\quad D_{\mathbf{w}}(\mathbf{x}, \textbf{highway}) = 0.7$ $\quad D_{\mathbf{w}}(\mathbf{x}, \textbf{coast}) = -1.5$

Intuition in other non-visual contexts, MIL, **h** ⇔ localization

- Text classification: example with recipe webpages (VISIIR)
  - **x** recipe text (steps of recipe), **h** recipe step, **y** recipe label
  - Lasagna recipe:



  - $\mathbf{h}^{-}_{pizza}$ (boil, water): **negative evidence for class pizza**
- Molecule, *e.g.* **x** DNA, **h** DNA region, **y** chemical property
  - $\mathbf{h}^{-}$ inhibition region in DNA for the chemical property

MANTRA: Optimization

- MANTRA Instantiation: define $(\mathbf{x}, \mathbf{y}, \mathbf{h})$, $\Psi(\mathbf{x}, \mathbf{y}, \mathbf{h})$, $\Delta(\mathbf{y}_i, \mathbf{y})$
- Instantiations: binary & multi-class classification, AP ranking

	Binary	**Multi-class**	**AP Ranking**
$\mathbf{x}$	bag (image/text/molecule)	bag (set of regions)	set of bags (of regions)
$\mathbf{y}$	± 1	$\{1, \ldots, K\}$	ranking matrix
$\mathbf{h}$	instance (region)	region	regions
$\Psi(\mathbf{x}, \mathbf{y}, \mathbf{h})$	$\mathbf{y} \cdot \phi(\mathbf{x}, \mathbf{h})$	$\{I(\mathbf{y} = 1)\Phi(\mathbf{x}, \mathbf{h}), ..$ $, I(\mathbf{y} = K)\Phi(\mathbf{x}, \mathbf{h})\}$	joint latent ranking feature map
$\Delta(\mathbf{y}_i, \mathbf{y})$	0/1 loss	0/1 loss	AP loss
LAI	exhaustive	exhaustive	exact and efficient

- Solve Inference $\max_y D_\mathbf{w}(\mathbf{x}_i, \mathbf{y})$ & LAI $\max_y \left[\Delta(\mathbf{y}_i, \mathbf{y}) + D_\mathbf{w}(\mathbf{x}_i, \mathbf{y})\right]$
  - Exhaustive for binary/multi-class classification
  - **Exact** and **efficient solutions** for ranking

MANTRA: Optimization

Latent structured AP ranking

- Latent feature map: $\Psi(\mathbf{x},\mathbf{y},\mathbf{h}) = \sum_{x_i \in \oplus} \sum_{x_j \in \ominus} y_{ij} [\Phi(x_i, h_{i,j}) - \Phi(x_j, h_{j,i})]$
  - $D(\mathbf{x}_i, \mathbf{y}) = \max_h \langle \mathbf{w}; \Psi(\mathbf{x},\mathbf{y},\mathbf{h}) \rangle + \min_h \langle \mathbf{w}; \Psi(\mathbf{x},\mathbf{y},\mathbf{h}) \rangle$
- **Lemma:** $D(\mathbf{x}_i, \mathbf{y}) = \sum_{x_i \in \oplus} \sum_{x_j \in \ominus} y_{ij} \left[\langle \mathbf{w}, \Phi^+_-(x_i) \rangle - \langle \mathbf{w}, \Phi^+_-(x_j) \rangle \right]$
  - $\langle \mathbf{w}, \Phi^+_-(x_i) \rangle = \max_{h \in \mathcal{H}_i} \langle \mathbf{w}, \Phi(x_i, h) \rangle + \min_{h \in \mathcal{H}_i} \langle \mathbf{w}, \Phi(x_i, h) \rangle$
  - $\sim$ Supervised problem with feature for each example $\mathbf{x}_i$: $\Phi^+_-(x_i)$

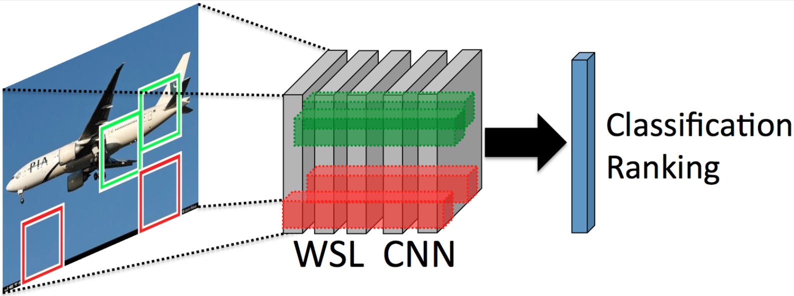
  ▷ **Elegant symmetrization due to the** `max+min` **scoring**
  ▷ **Decoupling optimization over y and h, ≠ [YJ09, BMJK15]**
- Inference: sort examples wrt $\langle \mathbf{w}, \Phi^+_-(x_i) \rangle$ scores
- LAI: $\sim$ supervised problem with $\Phi^+_-(x_i)$ feature for each $\mathbf{x}_i$, use [YFRJ07]

Outline

WELDON

Weakly Supervised Learning of Deep Convolutional Neural Networks

- MANTRA extension for training deep CNNs

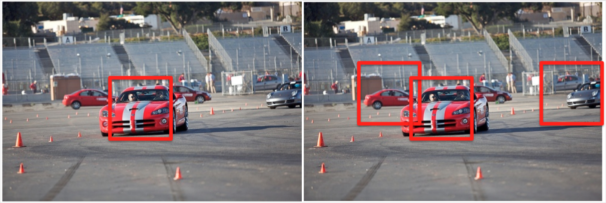


- learning $\Psi(\mathbf{x}, \mathbf{y})$: end-to-end WSL of deep CNNs with structured prediction
  - Incorporating multiple positive & negative evidence
  - Training deep CNNs with structured loss
  - Architectural choices ⇒ efficiency & robustness to over-fitting

WELDON: Model & Training

Region selection policy: k-max + k-min pooling

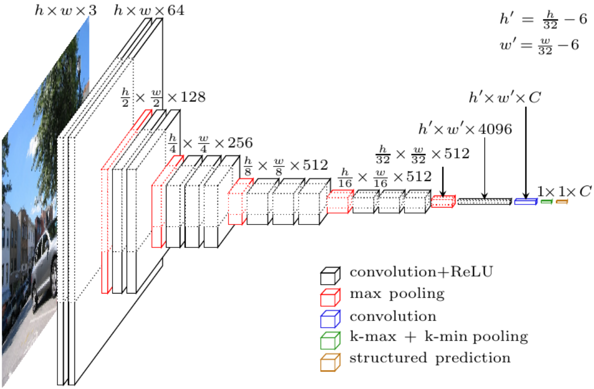
- Top-instances selection [LV15]: Σ k-max scores ⇒ convex
- Adding k-min (negative evidence): Σ k-min scores ⇒ concave
- Using more instances ⇒ robustness to outliers



Training: optimization for structured ranking

- MANTRA generalization for k-max + k-min: exact solutions
  - Inference: sorting wrt k-max + k-min scores
  - LAI: each example represented by k-max + k-min features

WELDON: WSL deep architecture

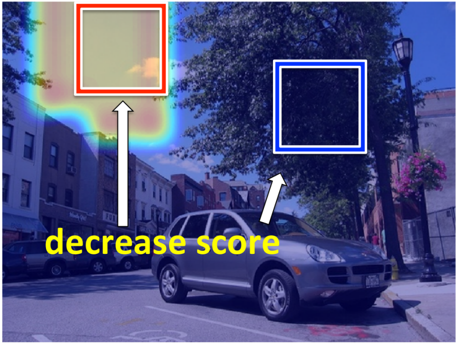


- Convolutional architecture
  - Efficient region feature computation
  - ImageNet transfer
- Fine-tuning
  ⇒ end-to-end training
- MATRA + top instances
  ⇒ k-max + k-min
- Structured ranking AP loss for k-max + k-min

WELDON Weakly Supervised Learning Insight

class is present: **Increase** score of selecting windows

class is absent: **Decrease** score of selecting windows



Outline

1. Context
2. Contributions
3. Results

Negative Evidence Models: Results

Multiple Instance Learning (MIL)

- MIL datasets, binary classification: image, text & molecule
- $\Psi(\mathbf{x}, \mathbf{y}, \mathbf{h})$: handcrafted features describing instances in bags
  - Image region descriptor, BoW for text passage *etc*

Method	Image	Musk	Text
mi-SVM	73.4	84.5	81.6
MI-SVM	75.5	81.7	80.3
LSVM	74.4	82.7	80
SyMIL	**80.2**	**89.2**	**84.8**
MICA	73.9	87.5	82.3
MIGraph	76.1	**90**	-
MI-CRF	78.5	86.7	-
GP-WDA	79	88.4	83.2
eMIL	77	85.3	82.7

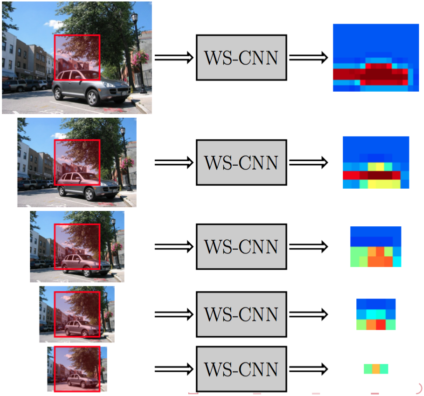
- `max+min` >> `max`
- ~ state-of-the-art results with more complex models (MI-CRF, MIGraph)

Negative Evidence Models: Visual Recognition Results

- $\Psi(\mathbf{x}, \mathbf{y}, \mathbf{h})$: deep features on regions
  MANTRA transfer (ImageNet, Places)
  WELDON fine-tuning (target dataset)
- Instantiations: Multi-class classification & ranking

Dataset	# ex	# class	Eval
VOC07	10k	20	AP
VOC12	10k	20	AP
15 Scene	5k	15	MC
MIT67	7k	67	MC
VOC12 act	4k	10	AP
COCO	120k	80	AP

- Multi-scale: 8 scales (Object Bank)



Negative Evidence Models: Visual Recognition Results

State-of-the-art results

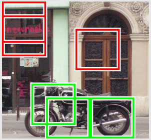
Multi-label (mAP)	VOC 2007	VOC 2012
VGG16	84.5	82.8
SPP net	82.4	
Deep WSL MIL		81.8
MANTRA	85.8	
WELDON	**90.2**	**88.5**
Multi-label (mAP)	**VOC12 Action**	**COCO**
VGG16	67.1	59.7
Deep WSL MIL		62.8
WELDON	**75.0**	**68.8**
Multi-class (acc)	**15 Scene**	**MIT67**
VGG16	91.2	69.9
MOP CNN		68.9
MANTRA	93.3	76.6
Negative parts		77.1
WELDON	**94.3**	**78.0**

Negative Evidence Models: Results

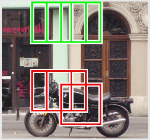
Impact of the different improvements

a) `max`	b) +k=3	c) `+min`	d) +AP	VOC07	VOC12 action
✓				83.6	53.5
✓	✓			86.3	62.6
✓		✓		87.5	68.4
✓		✓	✓	88.4	71.7
✓	✓	✓		87.8	69.8
✓	✓	✓	✓	**88.9**	**72.6**

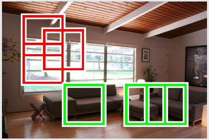
Detection results ??



Motorbike (1.1)



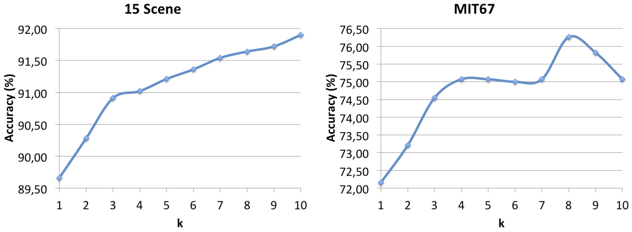
Sofa (-0.8)



Sofa (1.2)

Horse (-0.6)

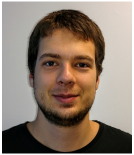
Negative Evidence Models: Visual Results



<u>Take-home message:</u> Contributions at different levels:

- Model: prediction function `max`+(k-)`min` > `max`
  - Using (k-)top-instances help, but selection needed
- Weakly supervised learning
  - AP ranking optimization: AP loss > Acc loss
- Deep CNN extension: learning $\Psi(\mathbf{x}, \mathbf{y})$

<u>Future Works:</u> Exploring other structured output predictions tasks, *e.g.* semantic segmentation

Thibaut Durand Nicolas Thome Matthieu Cord

MLIA Team (Patrick Gallinari)
Sorbonne Universités - UPMC Paris 6 - LIP6

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

Thibaut Durand, Nicolas Thome, and Matthieu Cord.
WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks.
In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016.

Thibaut Durand, Nicolas Thome, and Matthieu Cord.
MANTRA: Minimum Maximum LSSVM for Image Classification and Ranking.
In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

References I

Aseem Behl, Pritish Mohapatra, C. V. Jawahar, and M. Pawan Kumar, *Optimizing average precision using weakly supervised data*, IEEE Trans. Pattern Anal. Mach. Intell. **37** (2015), no. 12, 2545–2557.

Liang-Chieh Chen, Alexander G. Schwing, Alan L. Yuille, and Raquel Urtasun, *Learning deep structured models*, Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, 2015, pp. 1785–1794.

Trinh-Minh-Tri Do and Thierry Artières, *Regularized bundle methods for convex and non-convex risks*, JMLR (2012).

Weixin Li and Nuno Vasconcelos, *Multiple instance learning for soft bags via top instances*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

Pritish Mohapatra, C.V. Jawahar, and M. Pawan Kumar, *Efficient optimization for average precision svm*, NIPS, 2014.

S. N. Parizi, A. Vedaldi, A. Zisserman, and P. F. Felzenszwalb, *Automatic discovery and optimization of parts for image classification*, Proceedings of the International Conference on Learning Representations (ICLR), 2015.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, *Large margin methods for structured and interdependent output variables*, Journal of Machine Learning Research, 2005, pp. 1453–1484.

References II

Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims, *A support vector method for optimizing average precision*, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2007, pp. 271–278.

Chun-Nam Yu and T. Joachims, *Learning structural svms with latent variables*, ICML, 2009.