Computer Science

ICML 2019

University of Illinois at Chicago

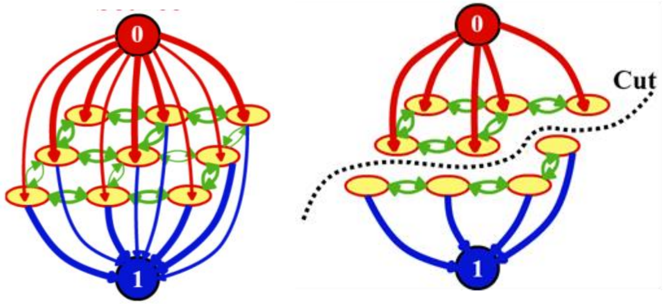
Active Learning for Probabilistic Structured Prediction of Cuts and Matchings

Sima Behpour, University of Pennsylvania
Anqi Liu, California Institute of Technology
Brian D. Ziebart, University of Illinois at Chicago

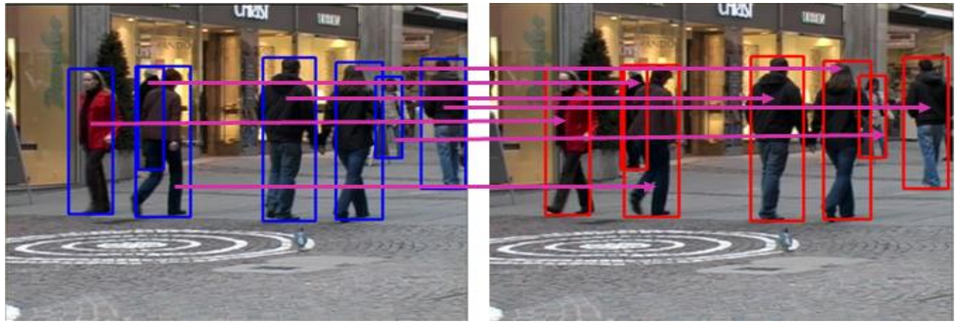
Motivation

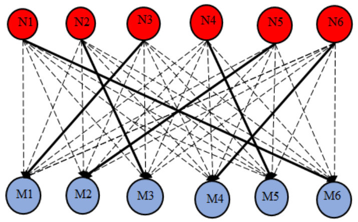
a) Multi-label Classification [Behpour et al. 2018]

Sea	0
Ship	0
Sheep	0
Wolf	0
Mountain	1
Person	1
Dog	1
Horse	1
Tree	1



b) Video Tracking





	M1	M2	M3	M4	M5	M6
N1	0	0	0	0	0	1
N2	0	0	1	0	0	0
N3	1	0	0	0	0	0
N4	0	0	0	0	1	0
N5	0	1	0	0	0	0
N6	0	0	0	1	0	0

Motivation

a) Multi-label Classification [Behpour et al. 2018]

Sea	0
Ship	0
Sheep	0
Wolf	0
Mountain	1
Person	1
Dog	1
Horse	1
Tree	1

Labeling can be

- Time consuming, e.g., document classification
- Expensive, e.g., medical decision (need doctors)
- Sometimes dangerous, e.g., landmine detection

b) Video Tracking

	M1	M2	M3	M4	M5	M6
N1	0	0	0	0	0	1
N2	0	0	1	0	0	0
N3	1	0	0	0	0	0
N4	0	0	0	0	1	0
N5	0	1	0	0	0	0
N6	0	0	0	1	0	0

Motivation

Active learning methods, like **uncertainty sampling**, combined with **probabilistic prediction techniques** [Lewis & Gale, 1994; Settles, 2012] have been successful.

Previous methods:

- **CRF**

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \prod_{n=1}^{N} p(\mathbf{y}_n | \mathbf{x}_n, \mathbf{w}) \qquad p(\mathbf{y}|\mathbf{x},\mathbf{w}) = \frac{\exp(\mathbf{w}^T \phi(\mathbf{x},\mathbf{y}))}{\sum_{\mathbf{y}} \exp(\mathbf{w}^T \phi(\mathbf{x},\mathbf{y}))}$$

  - Intractable

- **SSVM**

$$\min_{\boldsymbol{w}} \ \|\boldsymbol{w}\|^2 + C \sum_{n=1}^{\ell} \max_{y \in \mathcal{Y}} \left(0, \Delta(y_n, y) + \langle \boldsymbol{w}, \Psi(\boldsymbol{x}_n, y) \rangle - \langle \boldsymbol{w}, \Psi(\boldsymbol{x}_n, y_n) \rangle \right)$$

  - SVM Platts [Lambrou et al., 2012; Platt, 1999] ➔ Unreliable
  - Complication of Interpretation for multi-class

Our approach

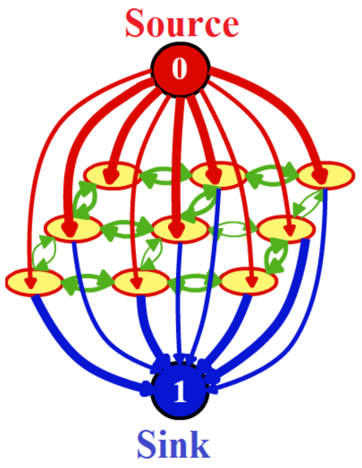
1- Leveraging Adversarial prediction methods [Behpour et al. 2018]:
- An Adversarial approximation of the training data labels, $\check{P}(\check{y}|x)$
- A predictor, $\hat{P}(\hat{y}|x)$, that minimizes the expected loss against the worst-case distribution chosen by the adversary.

$$\min_{\hat{P}(\hat{y}|x)} \max_{\check{P}(\check{y}|x)} \mathbb{E}_{\tilde{P}(x)\hat{P}(\hat{y}|x)\check{P}(\check{y}|x)}\left[\text{Loss}(\hat{Y},\check{Y})\right]$$

$$\mathbb{E}_{\tilde{P}(\mathbf{x})\check{P}(\check{\mathbf{y}}|\mathbf{x})}\left[\sum_i \phi_i(\check{Y}_i,\mathbf{X})\right] = \mathbb{E}_{\tilde{P}(\mathbf{x},\mathbf{y})}\left[\sum_i \phi_i(Y_i,\mathbf{X})\right]$$

and

$$\mathbb{E}_{\tilde{P}(\mathbf{x})\check{P}(\check{\mathbf{y}}|\mathbf{x})}\left[\sum_{i\neq j} \phi_{i,j}(\check{Y}_i,\check{Y}_j,\mathbf{X})\right] \geq \mathbb{E}_{\tilde{P}(\mathbf{x},\mathbf{y})}\left[\sum_{i\neq j} \phi_{i,j}(Y_i,Y_j,\mathbf{X})\right]$$



Our approach

2- Computing Mutual Information to measure reduction in uncertainty [Guo and Greiner 2007].

The mutual information of two discrete random variable a and b:
(the amount of the information which is held between a and b)

$$I(Y_a;Y_b) = \sum_{y_a \in A} \sum_{y_b \in B} p(y_a, y_b) \log \left(\frac{p(y_a, y_b)}{p(y_a)p(y_b)} \right)$$

$$\boxed{I(y_a, y_b) = H(y_a) + H(y_b) - H(y_a, y_b)}$$

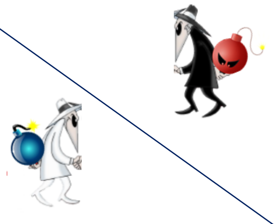
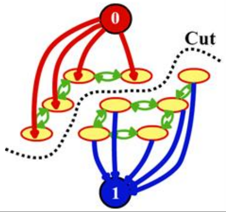
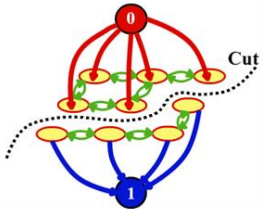
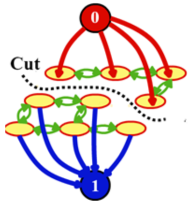
Marginal entropy of y_a

Marginal entropy of y_b

Joint entropy of y_a and y_b

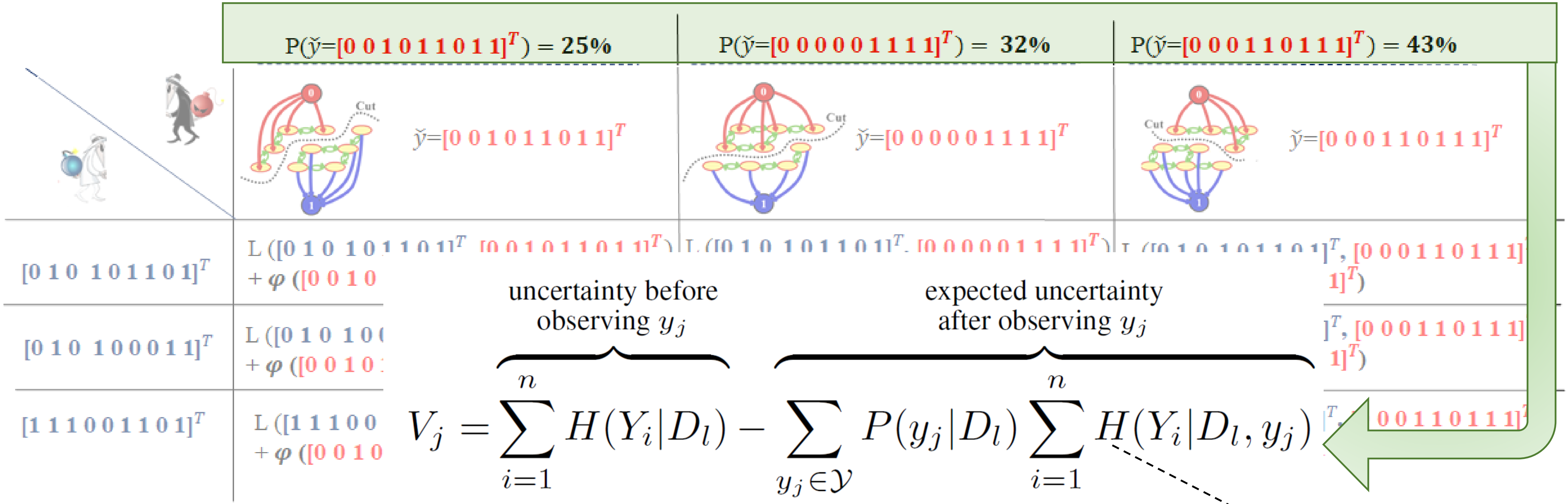
Game Matrix for Multi- label prediction

y = [Sea, Ship, Sheep, Horse, Dog, Person, Mountain, Wolf, Tree]

	$P(\check{y}=[0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1]^T) = 25\%$	$P(\check{y}=[0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]^T) = 32\%$	$P(\check{y}=[0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]^T) = 43\%$
	 $\check{y}=[0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1]^T$	 $\check{y}=[0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]^T$	 $\check{y}=[0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]^T$
$[0\ 1\ 0\ \ 1\ 0\ 1\ 1\ 0\ 1]^T$	$L([0\ 1\ 0\ \ 1\ 0\ 1\ 1\ 0\ 1]^T, [0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1]^T) + \varphi([0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1]^T)$	$L([0\ 1\ 0\ \ 1\ 0\ 1\ 1\ 0\ 1]^T, [0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]^T) + \varphi([0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]^T)$	$L([0\ 1\ 0\ \ 1\ 0\ 1\ 1\ 0\ 1]^T, [0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]^T) + \varphi([0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]^T)$
$[0\ 1\ 0\ \ 1\ 0\ 0\ 0\ 1\ 1]^T$	$L([0\ 1\ 0\ \ 1\ 0\ 0\ 0\ 1\ 1]^T, [0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1]^T) + \varphi([0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1]^T)$	$L([0\ 1\ 0\ \ 1\ 0\ 0\ 0\ 1\ 1]^T, [0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]^T) + \varphi([0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]^T)$	$L([0\ 1\ 0\ \ 1\ 0\ 0\ 0\ 1\ 1]^T, [0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]^T) + \varphi([0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]^T)$
$[1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1]^T$	$L([1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1]^T, [0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1]^T) + \varphi([0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1]^T)$	$L([1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1]^T, [0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]^T) + \varphi([0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]^T)$	$L([1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1]^T, [0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]^T) + \varphi([0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]^T)$

Sample selection strategy

The total expected reduction in uncertainty over all variables, $Y_1, \ldots, Y_n$, from Observing a particular variable Y_j



$$V_j = \overbrace{\sum_{i=1}^{n} H(Y_i|D_l)}^{\text{uncertainty before observing } y_j} - \overbrace{\sum_{y_j \in \mathcal{Y}} P(y_j|D_l) \sum_{i=1}^{n} H(Y_i|D_l, y_j)}^{\text{expected uncertainty after observing } y_j}$$

$$= \sum_{i=1}^{n} I(Y_i; Y_j|D_l).$$

Marginal entropy

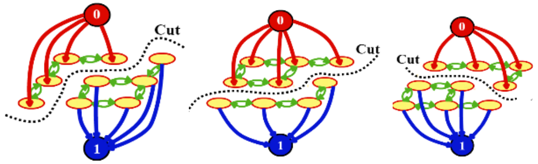
Active Learning for Cuts

Train a model

$\phi_i, \phi_{i,j}$

Test the model

$$\operatorname*{argmin}_{\check{y}} \sum_{\{i:\check{y}_i=1\}} \left(-\frac{1}{n}\sum_{\hat{y}} \hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \cdot \left[I(\hat{y}_i=0) - \boldsymbol{\psi}_i(1,\mathbf{x})\right]\right) + \sum_{\{i:\check{y}_i=0\}} \left(-\frac{1}{n}\sum_{\hat{y}} \hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \cdot \left[I(\hat{y}_i=1) - \boldsymbol{\psi}_i(0,\mathbf{x})\right]\right) + \sum_{i\neq j} I(y_i \neq y_j) - \boldsymbol{\theta}_{i,j}\boldsymbol{\delta}_{i,j}(y_i,y_j,\mathbf{x}).$$



Analyze unlabeled data pool

Unlabeled data pool

Labeled data pool

$$V_j = \overbrace{\sum_{i=1}^{n} H(Y_i|D_l)}^{\text{uncertainty before observing } y_j} - \overbrace{\sum_{y_j \in \mathcal{Y}} P(y_j|D_l) \sum_{i=1}^{n} H(Y_i|D_l, y_j)}^{\text{expected uncertainty after observing } y_j}$$
$$= \sum_{i=1}^{n} I(Y_i; Y_j|D_l).$$

Solicit the sample with the **highest** V_j

Return the sample
if there is any **unannotated** label.

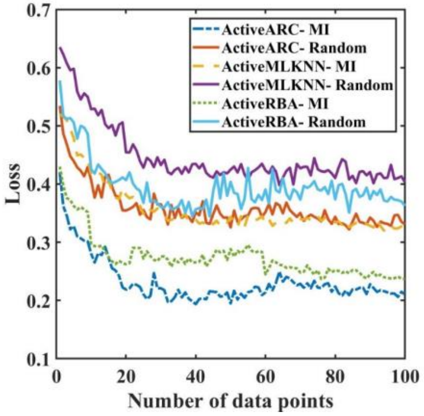
Y=[? 1 ? ? ? ? ? ? ? ?]

Add/ update the sample

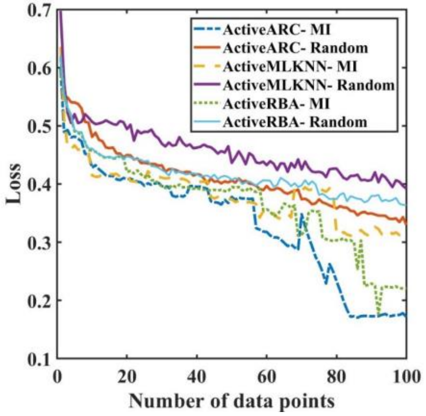
Y=[? 1 ? ? ? ? ? ? ? ?]

Multi-label Experiments

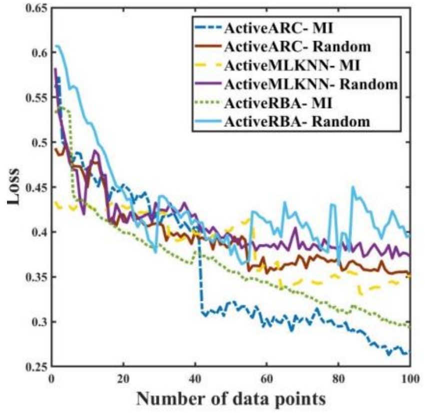
a) Bibtex



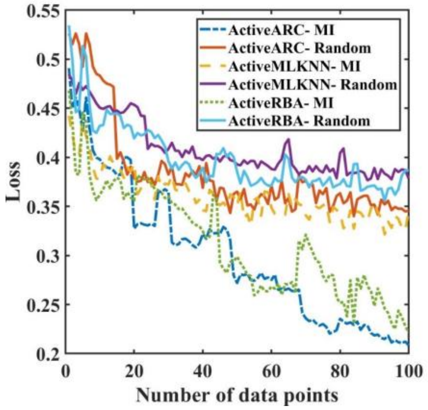
b) Bookmarks



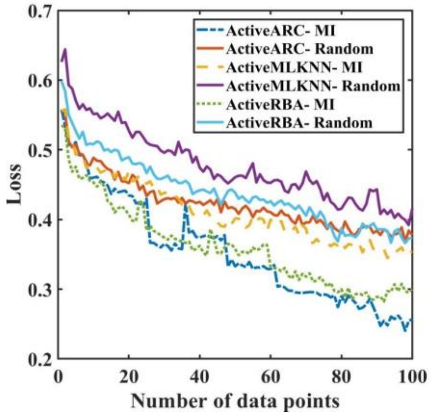
c) CAL500



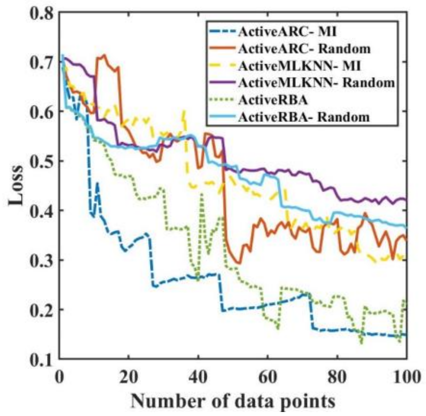
d) Corel5K



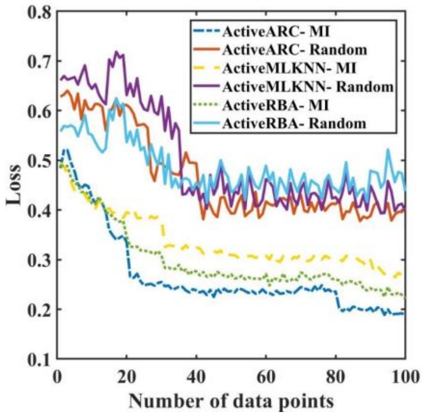
e) Enron



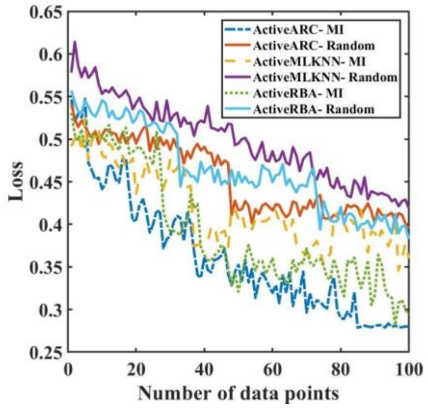
f) NUS-WIDE



g) TMC2007

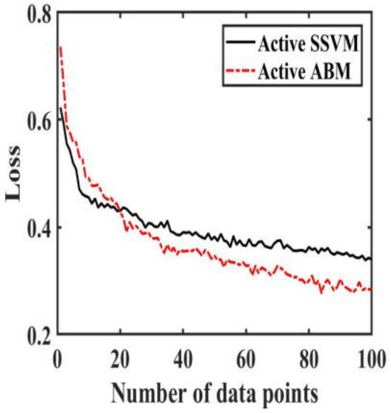


h) Yeast

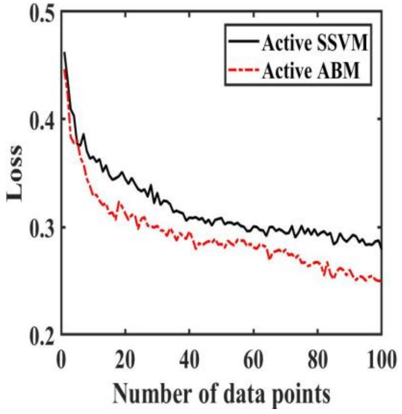


Tracking Experiments

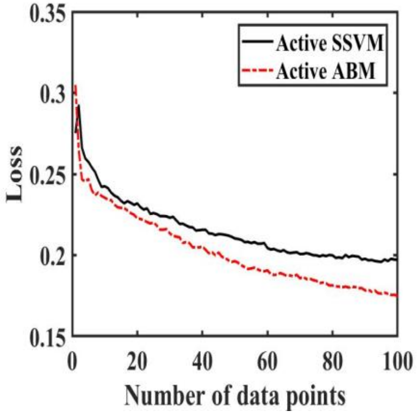
a) ETH-BAHNHOF



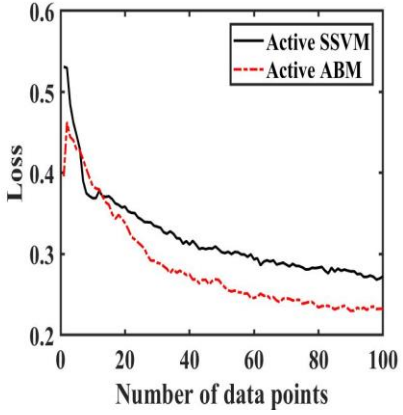
b) TUD-CAMPUS



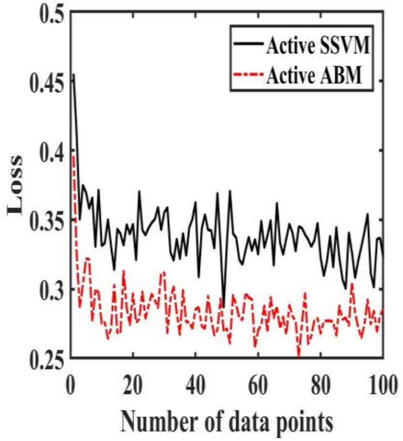
c) TUD-STADTMITTE



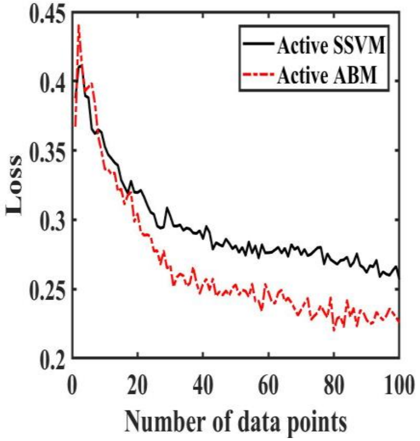
d) ETH-SUN



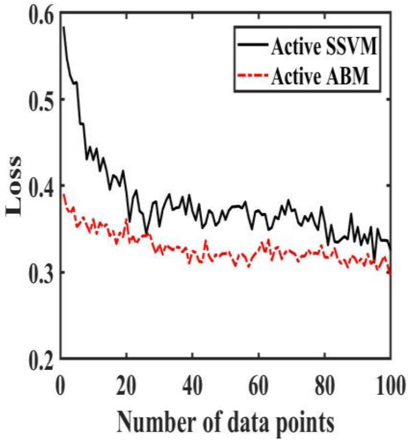
e) BAHNHOF-PEDCROSS2



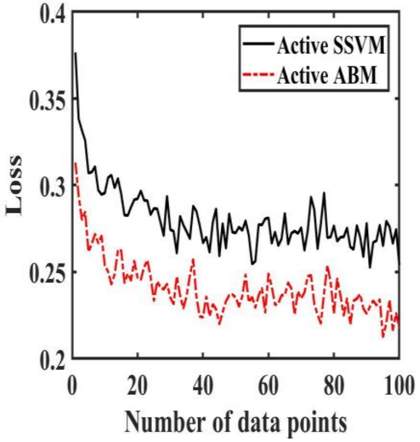
f) CAMPUS-STAD



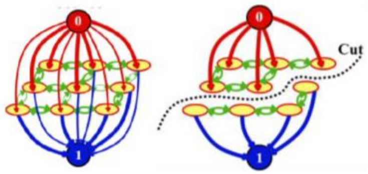
g) SUN-PEDCROSS2



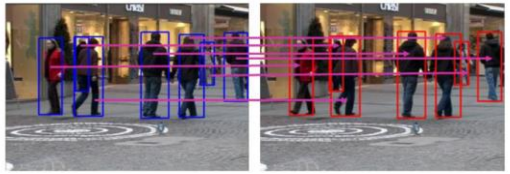
h) BAHNHOF-SUN



Conclusion



$$\min_{\hat{P}(\hat{y}|x)} \max_{\check{P}(\check{y}|x)} \mathbb{E}_{\tilde{P}(x)\hat{P}(\hat{y}|x)\check{P}(\check{y}|x)}[\text{Loss}(\hat{Y},\check{Y})]$$



$$\text{such that: } \mathbb{E}_{x\sim\tilde{P};\check{y}|x\sim\check{P}}\left[\phi(X,\check{Y})\right] = \tilde{\mathbf{c}}$$

Leveraging Adversarial Structured Predictions

- Adversarial Robust Cut
- Adversarial Bipartite Matching

Adversary probability distribution

correlations between unknown label variables

Useful in estimating
the value of information for different annotation solicitation decisions.

Better performance and **lower computational complexity**

Thank You!

Please visit our poster

at Pacific Ballroom #264