24th International Conference on Machine Learning, Oregon State University - Corvallis, OR, USA

On the Relation Between Multi-Instance Learning and Semi-Supervised Learning

Zhi-Hua Zhou and Jun-Ming Xu

{zhouzh, xujm}@lamda.nju.edu.cn

LAMDA Group,
National Key Laboratory for Novel Software Technology,
Nanjing University, China

MIL & SSL

Two different branches of machine learning:

- Multi-Instance Learning (MIL)
- Semi-Supervised Learning (SSL)

What is the relation between them?

What is implied by the relation?

Outline

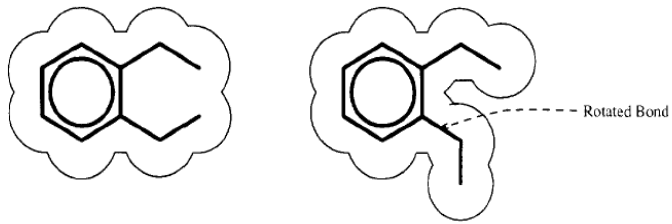
- Introduction
  - Multi-Instance Learning (MIL)
  - Semi-Supervised Learning (SSL)
- Our Work
- Experiments
- Conclusion and Discussion

Multi-Instance Learning

Originated from the research on drug activity prediction [Dietterich et al. AIJ97]

Drugs are small molecules working by binding to the target area

- For molecules qualified to make the drug, one of its shapes could tightly bind to the target area
- A molecule may have many alternative shapes



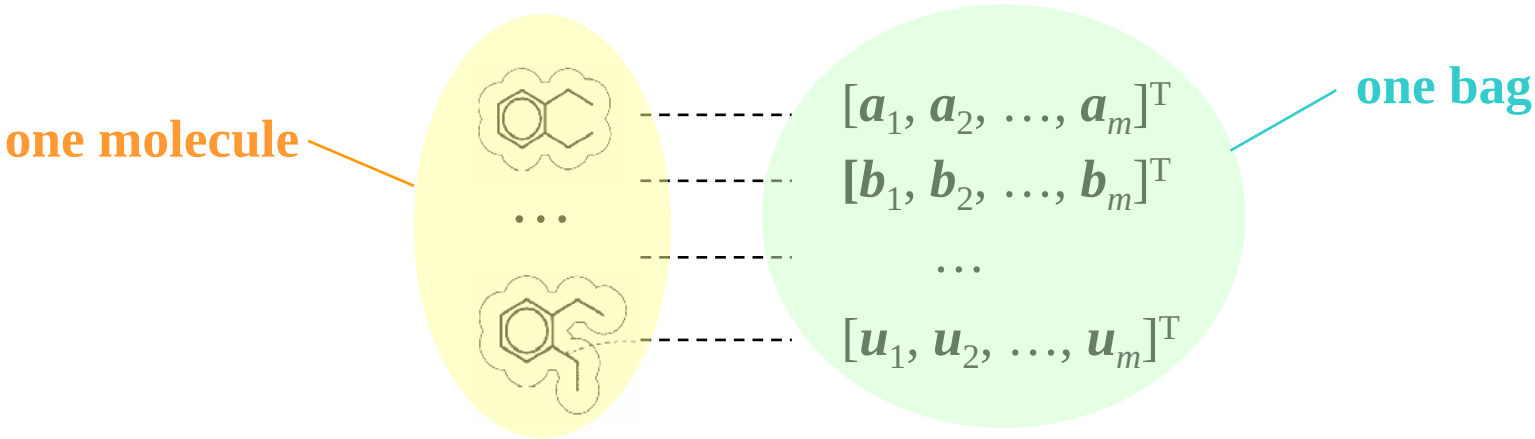
The difficulty:

Biochemists know that whether a molecule is qualified or not, but do not know which shape responses for the qualification

Figure reprinted from [Dietterich et al., AIJ97]

Multi-Instance Learning (con't)

Each shape can be represented by a feature vector, i.e., an instance



Thus, a molecule is a **bag** of instances

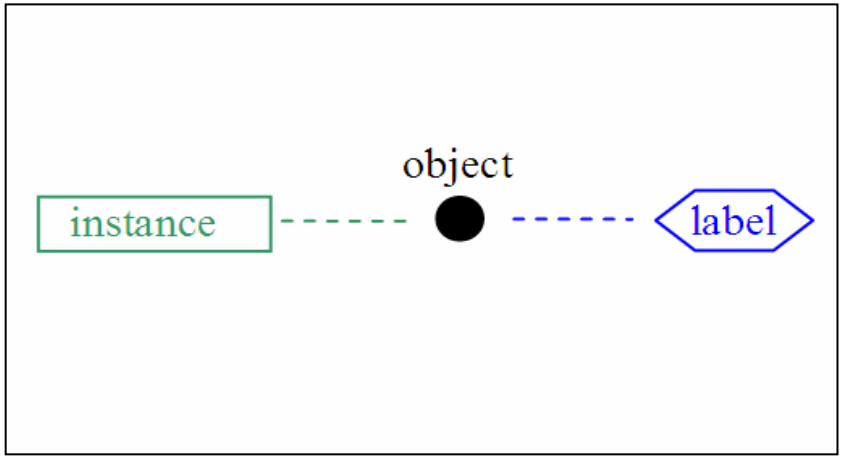
- A bag is positive if it contains at least one positive instance; otherwise it is negative
- The labels of the training bags are known
- The labels of the instances in the training bags are unknown

Multi-Instance Learning (con't)

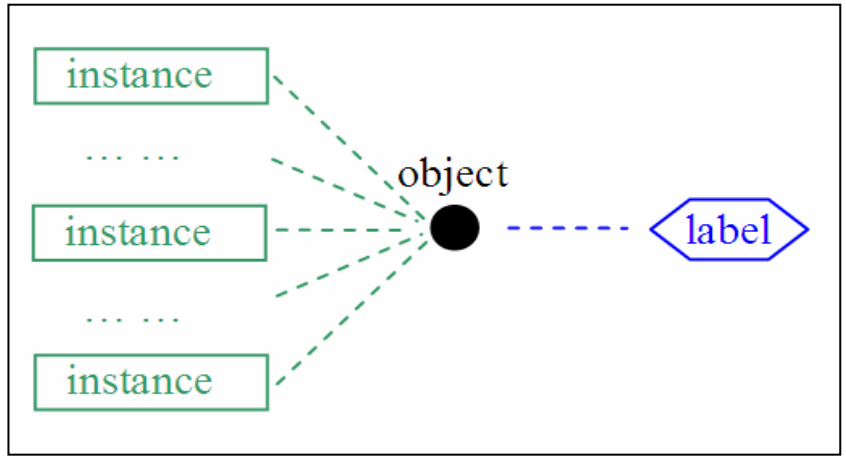
Formally:

Given a data set $\{(X_1, y_1), (X_2, y_2), \cdots, (X_m, y_m)\}$, where $X_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{i,n_i}\} \subseteq \mathcal{X}$ $(i \in \{1, \cdots, m\})$ is a bag and $y_i \in \{-1, +1\}$ is a class label, the goal is to predict the labels of unseen bags. X_i is a positive bag (thus $y_i = +1$) if there exists $g \in \{1, \cdots, n_i\}$, $\boldsymbol{x}_{ig}$ is positive. Yet the value of the index g is unknown.

Multi-Instance Learning (con't)



Traditional supervised learning



Multi-instance learning

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

Multi-Instance Learning (con't)

Many MIL algorithms:

- ✓ Diverse Density [Maron & Lozano-Perez, NIPS'97], EM-DD [Zhou & Goldman, NIPS'01]
- ✓ kNN algorithm: Citation-kNN [Wang & Zucker, ICML'00]
- ✓ Decision tree algorithms: RELIC [Ruffo, Thesis00], ID3-MI [Chevaleyre & Zucker, CanadianAI'01]
- ✓ Rule learning algorithm: RIPPER-MI [Chevaleyre & Zucker, CanadianAI'01]
- ✓ SVM algorithms: MI-SVM [Andrews et al., NIPS'02], mi-SVM [Andrews et al., NIPS'02], DD-SVM [Chen & Wang, JMLR04]
- ✓ Ensemble algorithms: MI-Ensemble [Zhou & Zhang, ECML'03], MI-Boosting [Xu & Frank, PAKDD'04]
- ✓ Logistic regression algorithm: MI-LR [Ray & Craven, ICML'05]
- ✓ … …

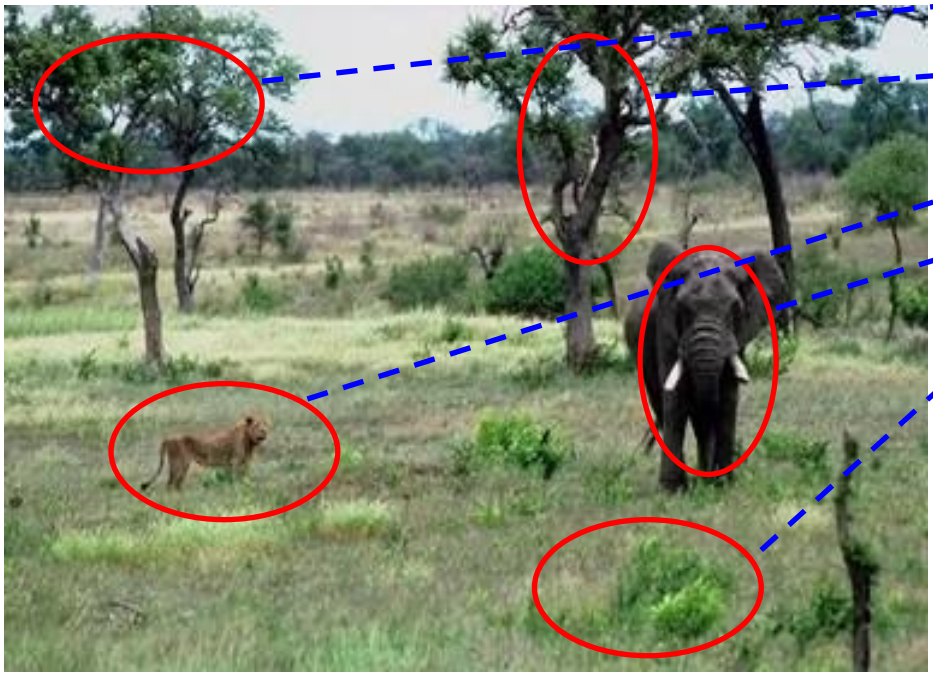
Multi-Instance Learning (con't)

Many Applications:

- ✓ Drug prediction [Dietterich et al., AIJ97]
- ✓ Image categorization [Maron & Ratan, ICML'98; Chen & Wang, JMLR04; Chen et al., PAMI06]
- ✓ Computer security [Ruffo, Thesis00]
- ✓ Web mining [Zhou et al., APIN05]
- ✓ Face detection [Viola et al., NIPS'05]
- ✓ … …

Multi-Instance Learning (con't)

Many tasks can be modeled as an MIL task



$[a_1, a_2, \ldots, a_m]^T$

$[b_1, b_2, \ldots, b_m]^T$

…

…

$[u_1, u_2, \ldots, u_m]^T$

image ⇨ bag

regions in the image ⇨ instances in the bag

Semi-Supervised Learning

In many practical applications, **unlabeled** training examples are readily available but labeled ones are fairly expansive to obtain because labeling the unlabeled examples requires human effort

The SSL task: To exploit unlabeled data to help improve the performance of learning with limited labeled data

Semi-Supervised Learning (con't)

Many SSL approaches:

- ✓ Generative model + EM

  Using a generative model for the classifier and employing EM to model the label estimation or parameter estimation process [Miller & Uyar, NIPS'96; Nigam et al., MLJ00; etc.]

- ✓ Regularization with unlabeled data

  Using unlabeled data to regularize the learning process in various ways [Belkin et al., COLT'01; Zhu et al., ICML'03; Zhou et al., NIPS'04; etc.]

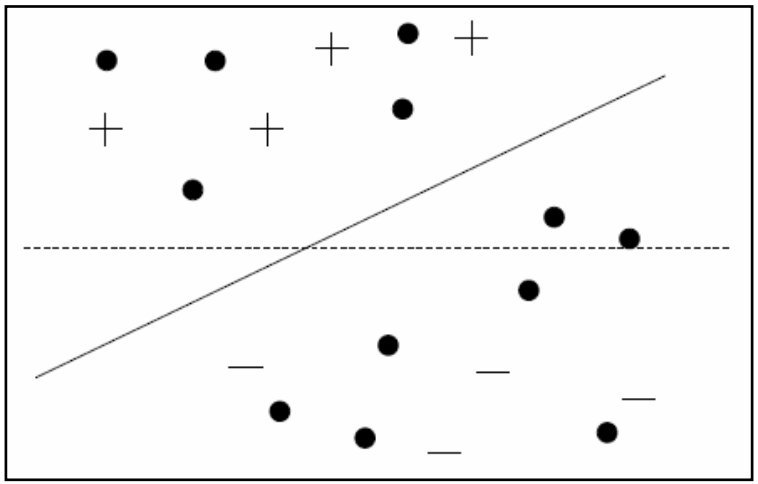
- ✓ Co-training

  Using two learners and letting them to label unlabeled instances for each other [Blum & Mitchell, COLT'98; Goldman & Zhou, ICML'00; Zhou & Li, IJCAI'05; etc.]

Semi-Supervised Learning (con't)

Semi-Supervised Support Vector Machines:

To maximize the margin on both labeled and unlabeled data



Earlier work:

- ✓ TSVM [Joachims, ICML'99]
- ✓ S^3VM [Bennett & Demiriz, NIPS'98]
- ✓ V^3SVM and CV^3SVM [Fung & Mangasarian, Techrep99]
- ✓ … …

Figure reprinted from [Joachims, ICML'99]

Semi-Supervised Learning (con't)

Some recent work on semi-supervised SVM:

- ✓ Multi-class Semi-supervised SVM based on SDP [Xu & Schuurmans, AAAI'05]
- ✓ Using branch-and-bound search to obtain global optimum on a small amount of data [Chapelle et al., NIPS'06]
- ✓ Using CCCP (Constrained Convex-Concave Procedure) to speed up semi-supervised SVMs [Collobert et al., ICML'06]
- ✓ … …

Multi-Instance Semi-Supervised Learning

MISSL (Multi-Instance Semi-Supervised Learning):

To exploit abundant unlabeled bags to help improve the performance of learning with a small number of labeled bags [Rahmani & Goldman, ICML'06]

Very different from our work although MISSL also involves both MIL and SSL

Notations

Assume there are p positive bags and q negative bags, $p + q = m$

The training bags can be organized as:

$$\{X_1^-, X_2^-, \cdots, X_q^-, X_{q+1}^+, \cdots, X_{q+p-1}^+, X_m^+\}$$

Put the instances bag-by-bag into an instance set:

$$\{\boldsymbol{x}_{11}, \cdots, \boldsymbol{x}_{1,n_1}, \boldsymbol{x}_{21}, \cdots, \boldsymbol{x}_{m1}, \cdots, \boldsymbol{x}_{m,n_m}\}$$

Re-index the set as $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$, where $T = \sum_{i=1}^{m} n_i$, then:

- the first $T_L = \sum_{i=1}^{q} n_i$ instances are from negative bags
- the remaining $T_U = \sum_{i=q+1}^{m} n_i$ instances are from positive bags
- the bag X_i 's instances are $\{\boldsymbol{x}_{s_i}, \cdots, \boldsymbol{x}_{e_i}\}$

$$s_i = \sum_{l=1}^{i-1} n_l + 1 \qquad e_i = \sum_{l=1}^{i} n_l = s_i + n_i - 1$$

A Reformulation of MIL Task

The definition of MIL [Dietterich et al., AIJ97] implies that:

- Negative bags contain only negative instances

  Thus, We can regard instances from negative bags as labeled negative examples

- Positive bags can contain positive as well as negative instances

  Thus, we can regard instances from positive bags as unlabeled examples with positive constraints

A Reformulation of MIL Task (con't)

An semi-supervised learning task:

Definition 1 *Given a set of labeled negative examples* $\{(\boldsymbol{x}_1, -1), (\boldsymbol{x}_2, -1), \cdots, (\boldsymbol{x}_{T_L}, -1)\}$ *and a set of unlabeled instances* $\{\boldsymbol{x}_{T_L+1}, \cdots, \boldsymbol{x}_T\}$, *to learn a function* $F^s : \mathcal{X} \to \{-1, +1\}$ *subject to: For* $i = q+1, \cdots, m$, *at least one instance in* $\{\boldsymbol{x}_{s_i}, \cdots, \boldsymbol{x}_{e_i}\}$ *is positive.*

The task can be solved by semi-supervised SVM algorithm:

MissSVM (Multi-instance learning by semi-supervised SVM)

- The main focus of the paper is not the proposal of a new algorithm (although we do propose a new algorithm)
- Instead of designing elaborate method, we try to use typical and simple SSL technique

The MissSVM Algorithm

The optimization problem for popular semi-supervised support vector machine:

$$\min_f \frac{1}{2}||f||_{\mathcal{H}}^2 + \lambda \sum_{t=1}^{T_L} H_1(y_t f(\boldsymbol{x}_t)) + \delta \sum_{t=T_L+1}^{T} D(f(\boldsymbol{x}_t))$$

where $H_1(z) = \max\{0, 1 - z\}$ is hinge loss

$$D(z) = \min\{H_1(z), H_1(-z)\}$$ [Bennett & Demiriz, NIPS'98]

Considering the positive constraints, the term should be added:

$$\sum_{i=q+1}^{m} H_1\left(\max_{t=s_i,\cdots,e_i} f(\boldsymbol{x}_t)\right)$$

The MissSVM Algorithm (con't)

Thus, the optimization problem can be written as:

$$\min_{f\in\mathcal{H},\eta,\theta,\varepsilon,\xi} \frac{1}{2}||f||_{\mathcal{H}}^2 + \lambda\eta'\mathbf{1} + \gamma\theta'\mathbf{1} + \delta\min(\boldsymbol{\varepsilon},\boldsymbol{\xi})'\mathbf{1}$$

$$\text{s.t.} \begin{cases} (-1)f(\boldsymbol{x}_t) + \eta_t \geqslant 1,\ \eta_t \geqslant 0,\ t = 1, 2, \cdots, T_L; \\ \max\limits_{t=s_i,\cdots,e_i} f(\boldsymbol{x}_t) + \theta_{i-q} \geqslant 1,\ \theta_{i-q} \geqslant 0, \\ \qquad\qquad i = q+1, \cdots, m; \\ f(\boldsymbol{x}_t) + \varepsilon_{t-T_L} \geqslant 1,\ \varepsilon_{t-T_L} \geqslant 0, \\ \qquad\qquad t = T_L+1, \cdots, T; \\ (-1)f(\boldsymbol{x}_t) + \xi_{t-T_L} \geqslant 1,\ \xi_{t-T_L} \geqslant 0, \\ \qquad\qquad t = T_L+1, \cdots, T. \end{cases}$$

- $\eta = [\eta_1, \cdots, \eta_{T_L}]'$ - slack variables for errors on instances from negative bags
- $\theta = [\theta_1, \cdots, \theta_p]'$ - slack variables for errors on positive bags
- $\varepsilon = [\varepsilon_1, \cdots, \varepsilon_{T_U}]'$ and $\xi = [\xi_1, \ldots, \xi_{T_U}]'$ - slack variables for errors on instances from positive bags
- λ, γ and δ - parameters

The MissSVM Algorithm (con't)

Let $\mathbf{K}$ denote a $T \times T$ kernel matrix and let $\mathbf{k}_t$ denote the t-th column:

$$\min_{\boldsymbol{\alpha},\eta,\theta,\varepsilon,\xi,b} \frac{1}{2}\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} + \lambda\eta'\mathbf{1} + \gamma\theta'\mathbf{1} + \delta\min(\boldsymbol{\varepsilon},\boldsymbol{\xi})'\mathbf{1}$$

$$\text{s.t.} \begin{cases} (-1)(\mathbf{k}'_t\boldsymbol{\alpha} + b) + \eta_t \geqslant 1,\ \eta_t \geqslant 0, \\ \qquad\qquad t = 1, 2, \cdots, T_L; \\ \max\limits_{t=s_i,\cdots,e_i} (\mathbf{k}'_t\boldsymbol{\alpha} + b) + \theta_{i-q} \geqslant 1,\ \theta_{i-q} \geqslant 0, \\ \qquad\qquad i = q+1, \cdots, m; \\ (\mathbf{k}'_t\boldsymbol{\alpha} + b) + \varepsilon_{t-T_L} \geqslant 1,\ \varepsilon_{t-T_L} \geqslant 0, \\ \qquad\qquad t = T_L+1, \cdots, T; \\ (-1)(\mathbf{k}'_t\boldsymbol{\alpha} + b) + \xi_{t-T_L} \geqslant 1,\ \xi_{t-T_L} \geqslant 0, \\ \qquad\qquad t = T_L+1, \cdots, T. \end{cases}$$

After replacing the gradients of the non-smooth *min* and *max* by their subgradients, CCCP (Constrained Convex-Concave Procedure) [Smola et al., AISTATS'05] can be used to solve this optimization problem

The MissSVM Algorithm (con't)

We use:

- The definition of the subgradient of $\min(\varepsilon, \xi)'1$ defined in [Fung & Mangasarian, Techrep99]
- The definition of the subgradient of $\max_{t=s_i,\cdots,e_i} \mathrm{k'}_t\alpha$ defined in [Cheung & Kwok, ICML'06]

Drug Activity Prediction

- Musk1:
  47 positive bags, 45 negative bags, 2 to 40 (5.15 in average) instances per bag
- Musk2:
  39 positive bags, 63 negative bags, 1 to 1,044 (64.49 in average) instances per bag

Table 1. Predictive accuracy (%) on the *Musk* data

Algorithm	*Musk1*	*Musk2*
MissSVM	87.6	80.0
MILES	86.3	87.7
MI-LR	86.7	87.0
MIBoosting	87.9	84.0
DD-SVM	85.8	91.3
mi-SVM	87.4	83.6
MI-SVM	77.9	84.3
RIPPER-MI	88.0	77.0
RELIC	83.7	87.3
Citation-*k*NN	92.4	86.3
Diverse Density	88.9	82.5
MULTINST	76.7	84.0
Iterated-discrim APR	92.4	89.2

Image Categorization

- 2,000 COREL images
- ROIs in image are regarded as instances, each is a 9-dim feature vector
- 20 categories each contains 100 images
- #(instances per bag) are in average 2.00 to 7.59 for different categories
- The same data and same experimental routine as in [Chen & Wang, JMLR04; Chen et al., PAMI06]

Table 3. Overall accuracy (%) on image categorization

Algorithm	*1000-Image*	*2000-Image*
MissSVM	78.0: [75.8,80.2]	65.2: [62.0,68.3]
MILES	82.6: [81.4,83.7]	68.7: [67.3,70.1]
DD-SVM	81.5: [78.5,84.5]	67.5: [66.1,68.9]
MI-SVM	74.7: [74.1,75.3]	54.6: [53.1,56.1]
$k$means-SVM	69.8: [67.9,71.7]	52.3: [51.6,52.9]

Web Index Page Recommendation

- 9 data sets with different positive/negative ratios, 32.2MB after compression
- 113 web index pages; the linked pages are regarded as instances
- 4 to 200 (in average 30.29) instances per bag, each represented by the 1st to 15th most frequent terms
- The same data and same experimental routine as in [Zhou et al., APIN05]

Table 5. Results averaged across the nine data sets

Algorithm	Precision	Recall	F-measure
MissSVM	0.627	0.838	0.690
Fretcit-kNN	0.739	0.741	0.728
r-Fretcit-kNN	0.727	0.720	0.704
TFIDF	0.679	0.620	0.591

The First Question

Question: What is the relation between MIL and SSL ?

Answer: MIL can be viewed as a special case of SSL by assuming i.i.d. instances

- Note that when putting the instances from bags into an instance set, i.i.d. instances were implicitly assumed
- Most previous MIL studies took this assumption explicitly or implicitly

The Second Question

Question: What is implied by the relation?

Answer (in fact some new questions):

Why shall we have a separate MIL area? Why not merge MIL into SSL?

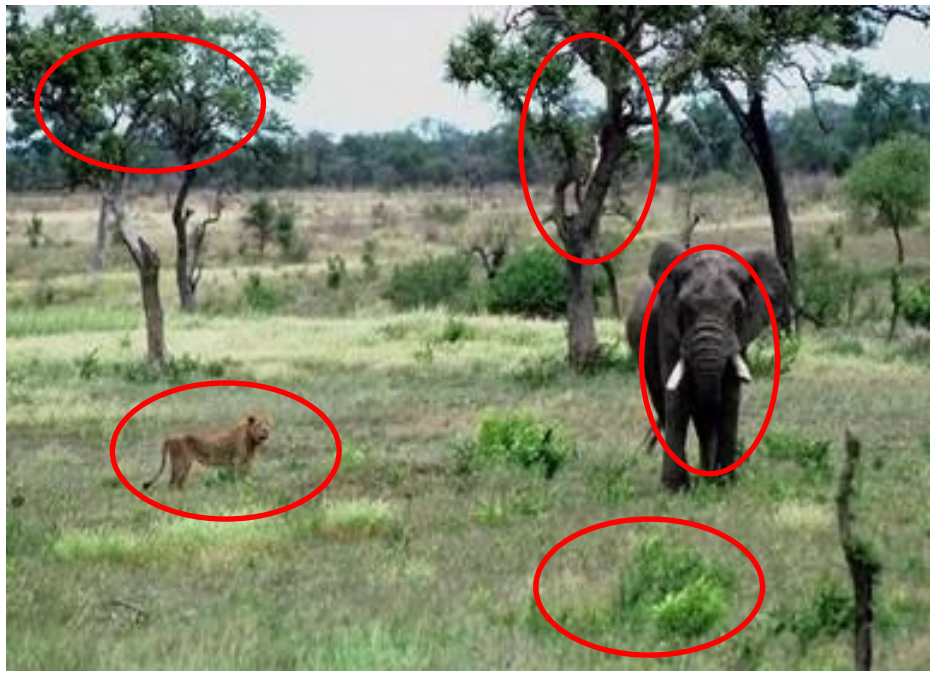
- Considering that MIL problems can be solved by SSL techniques, and by using better SSL techniques, better solution to MIL may be obtained. Why bother to design special MIL algorithms?

However, note that the premise of the above questions is: the assumption of i.i.d. instances is valid

The Second Question (con't)

A further question: Can we assume i.i.d. instances in MIL?

Answer: Usually it is not reasonable to make this assumption



We should not neglect the interaction between instances !!

Final Conclusion

If we assume i.i.d. instances, MIL would not be an important area since it can be viewed as a special case of SSL

Although most previous MIL research took this assumption, future MIL research should not assume this any more

Thanks!