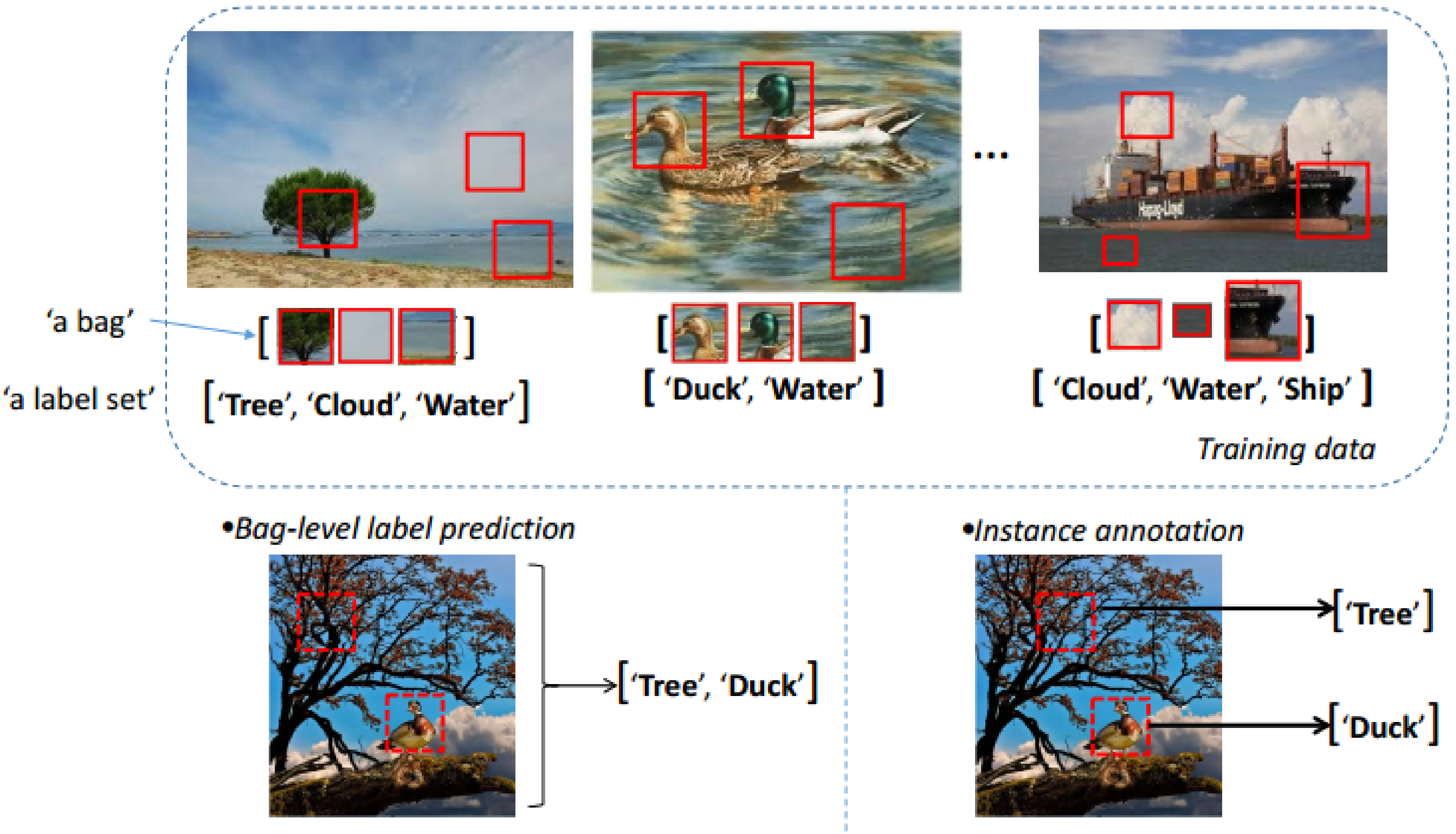
Multi-Instance Multi-label Learning in the Presence of Novel Class Instances

By Anh Pham, Raviv Raich, Xiaoli Fern

Presented By -
Akhil Gupta
Rohit Dayama
Suryansh Agnihotri
Sunit Pujari

Multi-Instance Multi-label learning (MIML)



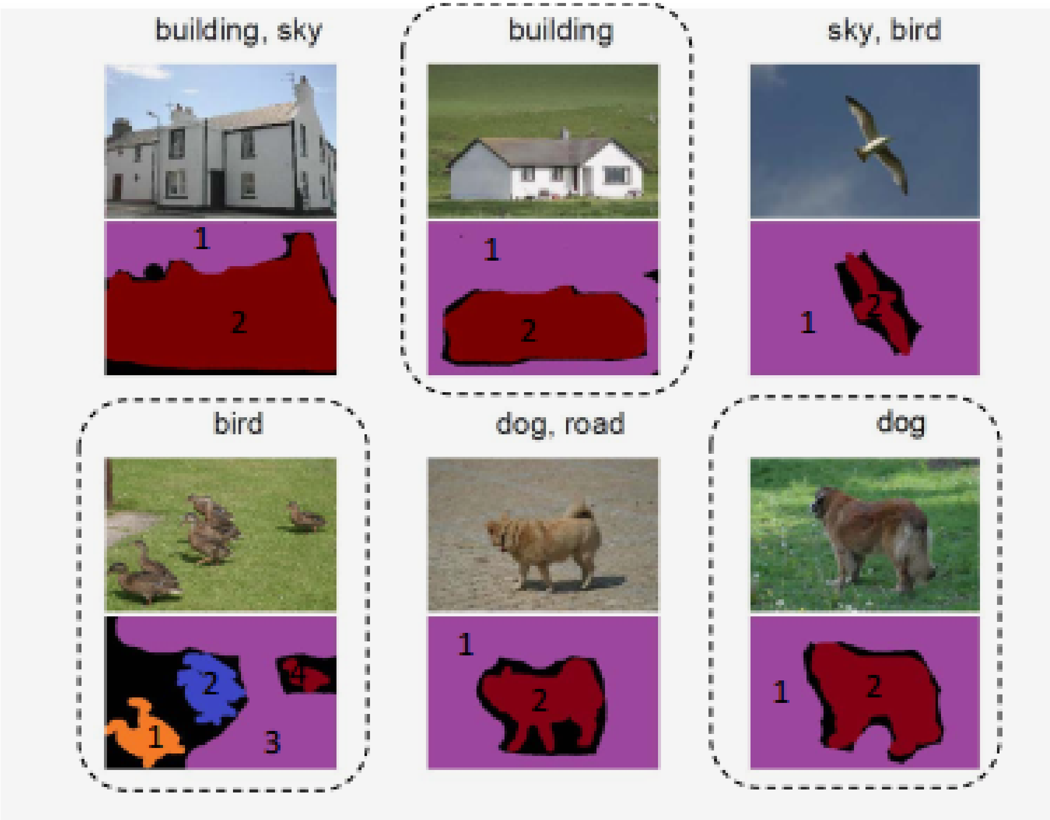
MIML learning with novel instances

Known Classes

- Building,Sky,Bird
- Dog,Road

Novel Classes

- Grass
- Others (void)



- The boxed images contain grass segments but have no label for grass

Problem formulation

- Training data: B bags, denoted by $\{(\mathbf{X}_b, \mathbf{Y}_b)\}_{b=1}^{B}$
- $\mathbf{X}_b$ is a set of n_b instances for the bth bag $\{\mathbf{x}_{b1}, \mathbf{x}_{b2}, \ldots, \mathbf{x}_{bnb}\}$, where $\mathbf{x}_{bi} \in \underline{X} = R^d$
- Each instance $\mathbf{x}_{bi}$ is associated with a label $y_{bi} \in \{0, 1, \ldots, C\}$,where C is the number of classes and 0 denotes novel class
- $\mathbf{Y}_b$ is the bag label for the bth bag and a subset of known labels Y = {1,2,..C}

Goals

- Instance annotation: map an instance in X to a label in {0,1,2,…,C}
- Novelty detection: map an instance in X to
- {0, Y}
- Bag label prediction: map a bag in 2^{X} to 2^{Y}

Related work

- **Novelty detection in SISL learning**
- Novel instances are not in training (Saligrama & Zhao, 2012)
- **Novelty detection in MIML learning**
- Novel instances are in training, however, their labels are not available
- A threshold approach can be use with other MIML instance annotation Algorithms
- ***Approach in Paper*** : a discriminative framework with a built-in novel classmodel

Graphical model

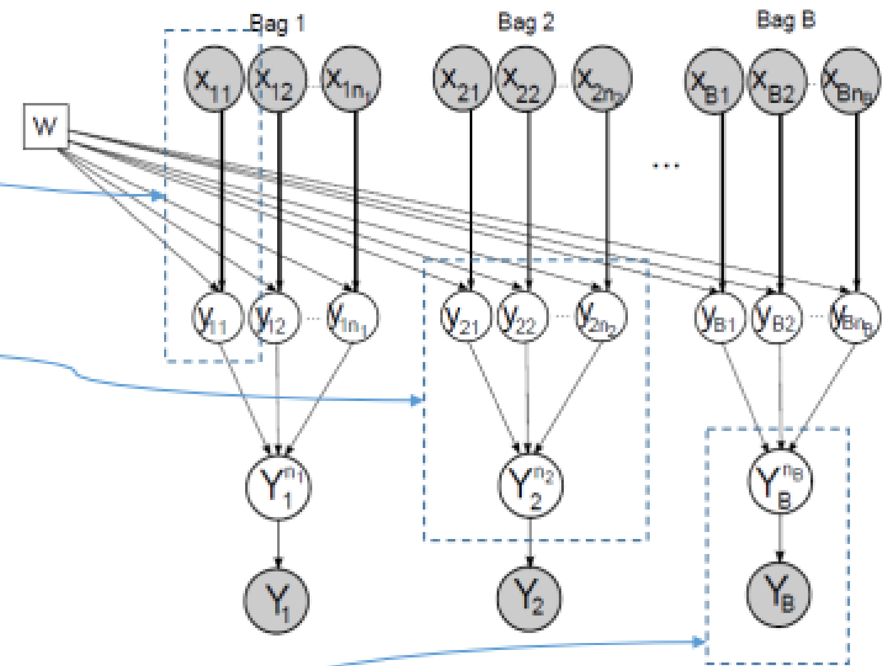
Multiclass Logistic regression

Union relation

y_{21} =`grass', $y_{22}$ =`dog' → $\mathbf{Y}_2^2$ ={`grass',`dog'}

Instance labels are 0,1,2, … , C where 0 is the novel class label

$$p(\mathbf{Y}_b|\mathbf{Y}_b^{n_b}) = I(\mathbf{Y}_b = \mathbf{Y}_b^{n_b}) + I(\mathbf{Y}_b \bigcup \{0\} = \mathbf{Y}_b^{n_b})$$



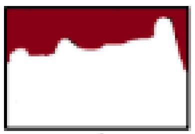
Graphical model for the problem

$\mathbf{Y}_b^{n_b}$ may include the novel class label 0 : $\mathbf{Y}_b^{n_b} \subseteq \{0,1,2, \dots, C\}$

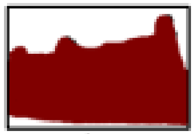
$\mathbf{Y}_b$ removes the novel label 0 from $\mathbf{Y}_b^{n_b}$: $\mathbf{Y}_b^{n_b} \rightarrow \mathbf{Y}_b \subseteq \{1, \dots, C\}$

Graphical Model (Example 1)

x_{21}



x_{22}



y_{21} = `sky'

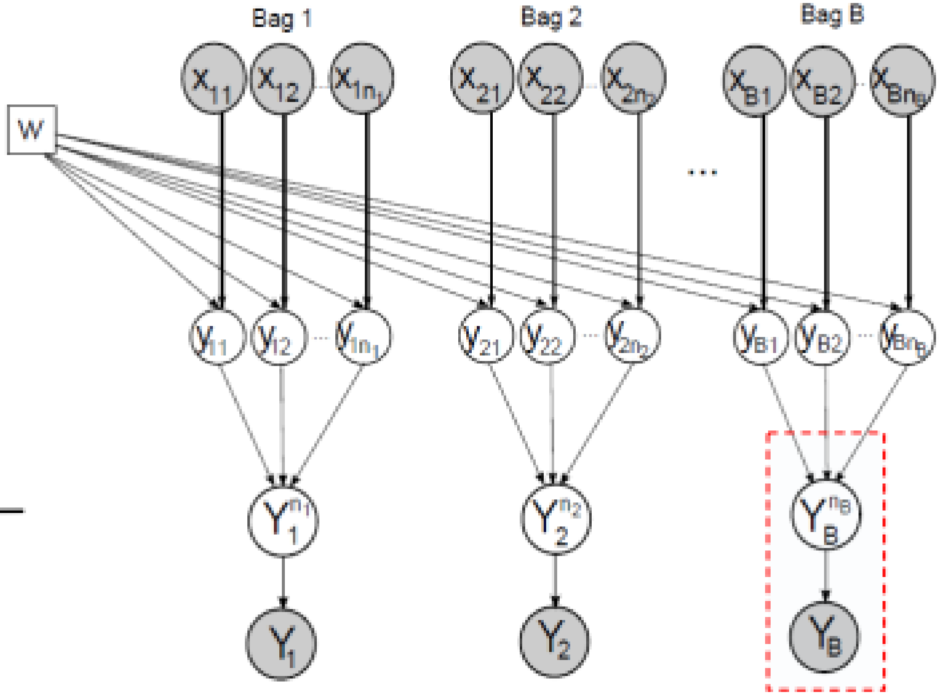
y_{22} = `building'

example

Y_2^2 = {`sky', ` building'}

since Y_2^2 does not include `grass' (0)

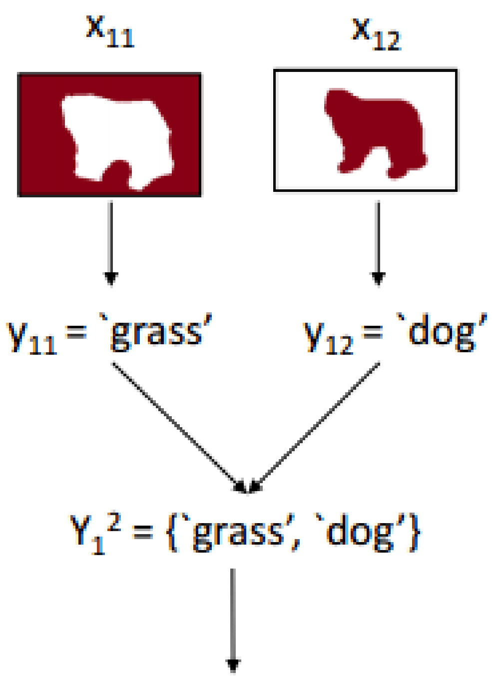
→ Y_2 = {`sky', ` building'}



Graphical model for the problem

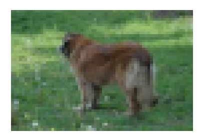
$$p(\mathbf{Y}_b|\mathbf{Y}_b^{n_b}) = I(\underbrace{\mathbf{Y}_b = \mathbf{Y}_b^{n_b}}_{1}) + I(\underbrace{\mathbf{Y}_b \bigcup \{0\} = \mathbf{Y}_b^{n_b}}_{0})$$

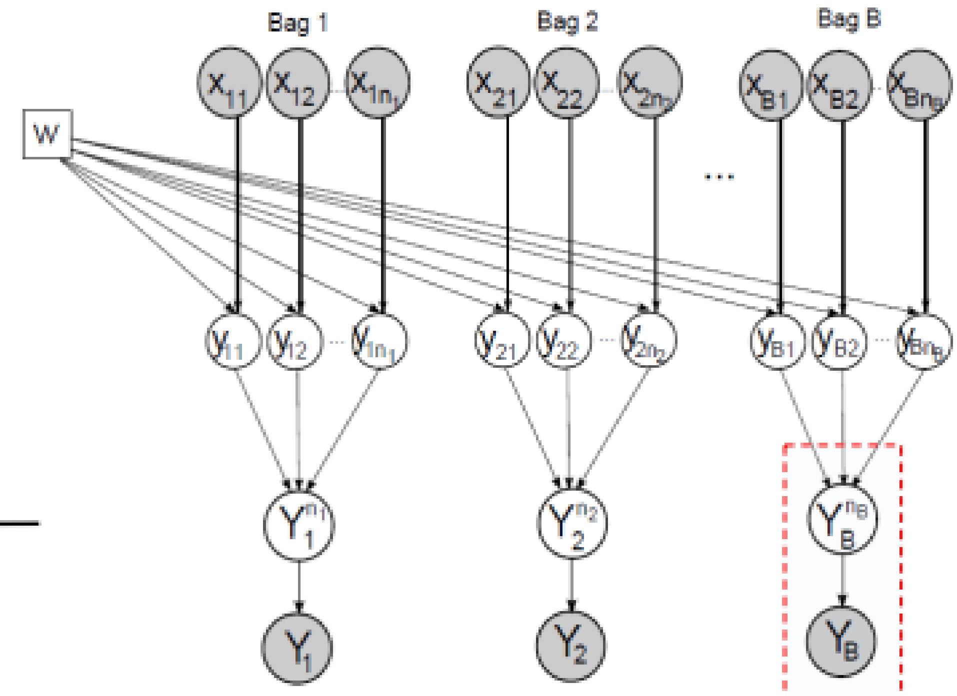
Graphical Model (Example 2)



since Y_1 cannot include the novel class `grass' (0)
→ Y_1 = {`dog' }

example





Graphical model for the problem

$$p(\mathbf{Y}_b|\mathbf{Y}_b^{n_b}) = \underbrace{I(\mathbf{Y}_b = \mathbf{Y}_b^{n_b})}_{0} + \underbrace{I(\mathbf{Y}_b \bigcup \{0\} = \mathbf{Y}_b^{n_b})}_{1}$$

Inference

- Maximum likelihood inference

$$p(\mathbf{Y}_D, \mathbf{X}_D|\mathbf{w}) = p(\mathbf{X}_D)\prod_{b=1}^{B} p(\mathbf{Y}_b|\mathbf{X}_b, \mathbf{w}) \text{ where}$$

$$p(\mathbf{Y}_b|\mathbf{X}_b,\mathbf{w}) = \sum_{y_{b1}=0}^{C}\cdots\sum_{y_{bn_b}=0}^{C}[\{I(\mathbf{Y}_b = \bigcup_{j=1}^{n_b} y_{bj}) + I(\mathbf{Y}_b \bigcup\{0\} = \bigcup_{j=1}^{n_b} y_{bj})\} \times \prod_{i=1}^{n_b} p(y_{bi}|\mathbf{x}_{bi},\mathbf{w})] \rightarrow O((C+1)^{n_b})$$

- Generalized Expectation Maximization
  - Surrogate function

$$g(\mathbf{w},\mathbf{w}') = E_{\mathbf{y}}[\log p(\mathbf{Y}_D,\mathbf{X}_D,\mathbf{y}|\mathbf{w})|\mathbf{Y}_D,\mathbf{X}_D,\mathbf{w}'] = \sum_{b=1}^{B}\sum_{i=1}^{n_b}[\sum_{c=0}^{C} p(y_{bi}=c|\mathbf{Y}_b,\mathbf{X}_b,\mathbf{w}')\mathbf{w}_c^T\mathbf{x}_{bi} - \log(\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_{bi}})] + \zeta,$$

  - E-step

Compute $p(y_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}^{(k)})$ → Still $O((C+1)^{n_b})$ for brute force marginalization

  - M-step

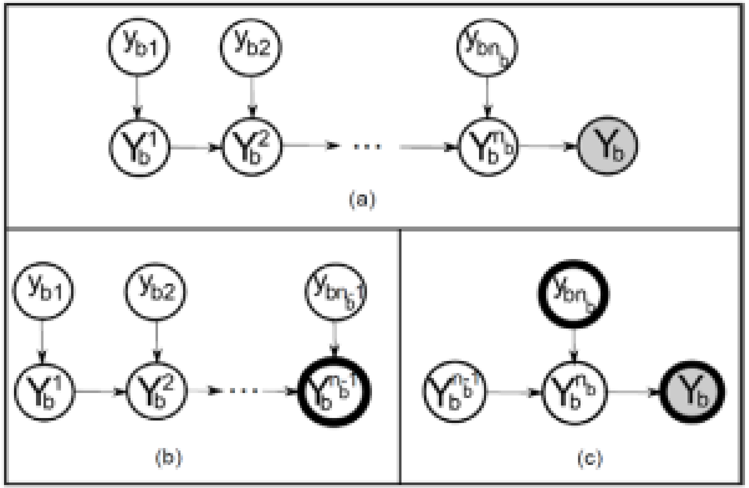
Find $\mathbf{w}^{(k+1)}$ such that $g(\mathbf{w}^{(k+1)}, \mathbf{w}^{(k)}) \geq g(\mathbf{w}^{(k)}, \mathbf{w}^{(k)})$

E-step: Compute $p(y_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}^{(k)})$

- Conditional rule:

$$p(y_{bi} = c|\mathbf{Y}_b = \mathbf{L}, \mathbf{X}_b, \mathbf{w}) = \frac{p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})}{\sum_{c \in \mathbf{L} \cup \{0\}} p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})}$$

- Compute $p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ for c in $\{\mathbf{Y}_b \cup \{0\}\}$



→ ***Linear*** computation time w.r.t. n_b

- Introduce a partial bag label

$$\mathbf{Y}_b^i = \bigcup_{k=1}^{i} y_{bk} \Rightarrow \mathbf{Y}_b^i = \mathbf{Y}_b^{i-1} \cup \{y_{bi}\}$$

Recall: $p(\mathbf{Y}_b|\mathbf{Y}_b^{n_b}) = I(\mathbf{Y}_b = \mathbf{Y}_b^{n_b}) + I(\mathbf{Y}_b \bigcup \{0\} = \mathbf{Y}_b^{n_b})$

- Allows for a recursive computation as follows
  - Compute $p(\mathbf{Y}_b^1)$
  - Compute $p(\mathbf{Y}_b^2)$
  - ....
  - Compute $p(\mathbf{Y}_b^{n_b-1})$
  - Finally, from $p(\mathbf{Y}_b^{n_b-1})$ and $p(y_{bn_b})$ → $p(y_{bn_b}, \mathbf{Y}_b)$

M Step

We apply gradient ascent to maximize g(w, w') w.r.t. w as follows:

$$\mathbf{w}_c^{(k+1)} = \mathbf{w}_c^{(k)} + \frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c}\bigg|_{\mathbf{w}=\mathbf{w}^{(k)}} \times \eta$$

where the gradient w.r.t. $\mathbf{w}_c$, for all $c \in \{0, 1, 2, \ldots, C\}$, $\frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c}$, is

$$\sum_{b=1}^{B} \sum_{i=1}^{n_b} [p(y_{bi} = c | \mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}^{(k)}) \mathbf{x}_{bi} - \frac{e^{\mathbf{w}_c^T \mathbf{x}_{bi}} \mathbf{x}_{bi}}{\sum_{c=0}^{C} e^{\mathbf{w}_c^T \mathbf{x}_{bi}}}]$$

Predictions

1. Instance Annotation

$$\hat{y}_{ti} = \arg\max_{0 \leq k \leq C} \mathbf{w}_k^T \mathbf{x}_{ti}.$$

2. Bag Label Prediction

$$\hat{\mathbf{Y}}_t = (\bigcup_{i=1}^{n_t} \hat{y}_{ti}) \backslash \{0\}$$

3. Novelty Detection

$$p(y_{ti} = 0 | \mathbf{x}_{ti}, \mathbf{w}) \geq \theta$$

Experimental Result

EM Iterations	Mean Accuracy
2	0.33
10	0.36
20	0.41
50	0.48
100	0.58
200	0.66