CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Large Loss Matters in Weakly Supervised Multi-Label Classification

Youngwook Kim, Jae Myung Kim, Zeynep Akata, Jungwoo Lee

SEOUL NATIONAL UNIVERSITY

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

MAX-PLANCK-GESELLSCHAFT

HodooAi

Multi-label classification

- Image can contain **multiple** categories
- Ground truth : Multi-hot vector
- It is gaining attention recently.
- Labelling cost is very expensive!

person
horse
cat
dog
truck

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Weakly supervised learning approach

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

"Weakly supervised multi-label classification" (WSML)

Partial label : Only small portion of full label is annotated per image (e.g. 10%)

- CVPR 2019, "Learning a Deep ConvNet for Multi-label Classification with Partial Labels"
- CVPR 2020, "Interactive Multi-Label CNN Learning with Partial Labels"
- NeurIPS 2020, "Exploiting weakly supervised visual patterns to learn from partial annotations"
- CVPR 2021, "Multi-Label Learning from Single Positive Labels"
- AAAI 2022, "Structured Semantic Transfer for Multi-Label Recognition with Partial Labels"
- AAAI 2022, "Semantic-Aware Representation Blending for Multi-Label Image Recognition with Partial Labels"

	[a]	[b]	[c]
person	1	1	1
horse	1		
cat	0		
dog	0	0	
truck	0		

[a] : full label / [b],[c] : partial label

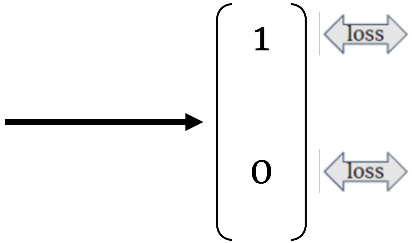
Learning with partial labels

Q. How to train the model with **incomplete labels**?

A1. Train the model using observed labels

+ Bootstrapping [CVPR 2019]
  Modeling label/image similarity from other images [CVPR 2020, NeurIPS 2020, AAAI 2022]
  Alternatively train image classifier and label estimator [CVPR 2021]

person	1
horse	?
cat	?
dog	0
truck	?



Limitation : Heavy, complex optimization process

Learning with partial labels

A2. Assume unobserved labels as negative (AN)

∵ Majorities of labels are negative in a multi-label setting [Ridnik et al, 2021]

person	1
horse	?
cat	?
dog	0
truck	?

→

AN target		
1	loss	
0	loss	False negative!
0	loss	True negative
0	loss	
0	loss	True negative

AN target

Limitation : Label noise produced

Learning with partial labels

A2. Assume unobserved labels as negative (AN)

∵ Majorities of labels are negative in a multi-label setting [Ridnik et al, 2021]

		AN target		
person	1	1	loss	
horse	?	0	loss	False negative!
cat	?	0	loss	True negative
dog	0	0	loss	
truck	?	0	loss	True negative

AN target

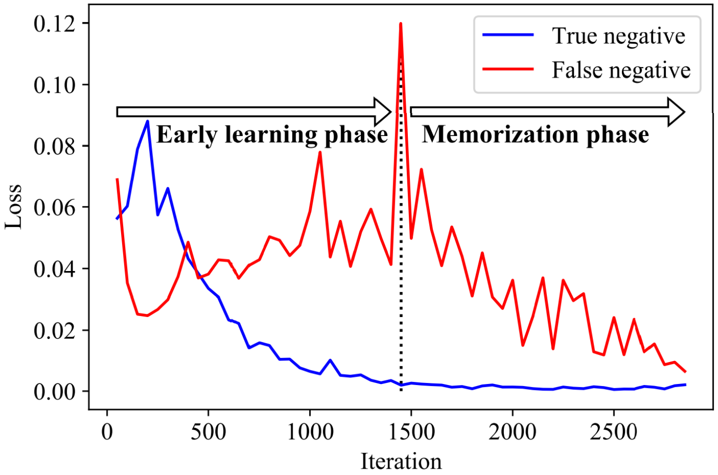
Limitation : Label noise produced

=> Look at the WSML problem from the perspective of **noisy label learning!**

Our key observation

When training a model with noisy AN target,
the model first fits into **clean label**
and then gradually fits into **noisy label**!

a.k.a. “Memorization effect” [Arpit et al., 2017]

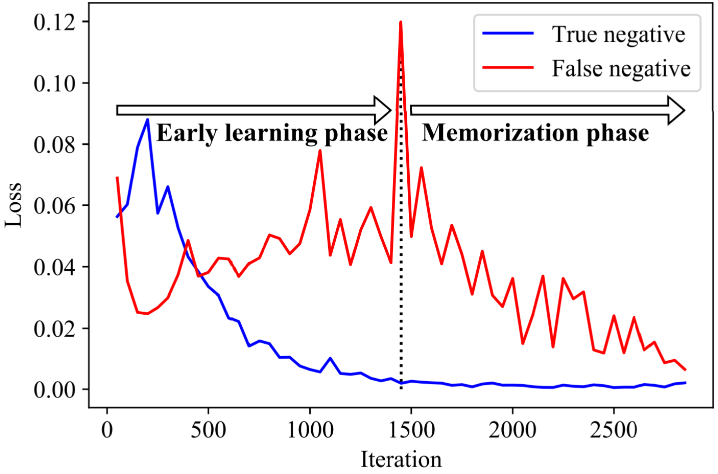


Highest loss phase	Pascal VOC (%) TP	TN	FN	MS COCO (%) TP	TN	FN
Warmup	**88.3**	**90.7**	23.8	**64.0**	**82.6**	17.3
Regular	11.7	9.3	**72.2**	36.0	17.4	**82.7**

Table 1. **Distribution of the highest loss occurrence.** For each label, we first draw the loss plot in the training process. We then record whether the highest loss occurred in the warmup phase (epoch 1) or in the regular phase (after epoch 1). TP, TN, FN refers to true positive, true negative, and false negative, respectively.

Our key observation

Based on memorization effect,
we can discriminate whether a specific sample is noisy
with its **loss value** during training! [Han et al., 2018]

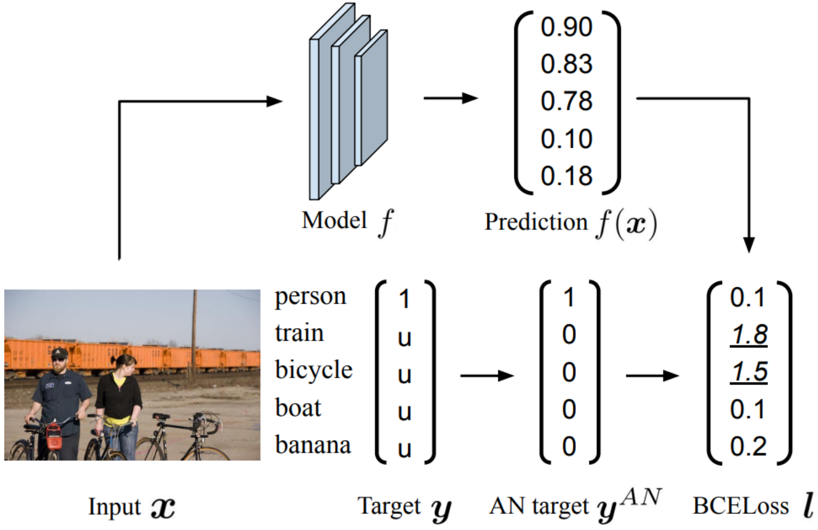


Highest loss phase	Pascal VOC (%)			MS COCO (%)		
	TP	TN	FN	TP	TN	FN
Warmup	**88.3**	**90.7**	23.8	**64.0**	**82.6**	17.3
Regular	11.7	9.3	**72.2**	36.0	17.4	**82.7**

Table 1. **Distribution of the highest loss occurrence.** For each label, we first draw the loss plot in the training process. We then record whether the highest loss occurred in the warmup phase (epoch 1) or in the regular phase (after epoch 1). TP, TN, FN refers to true positive, true negative, and false negative, respectively.

=> <u>**Reject or correct large loss samples during training!**</u>

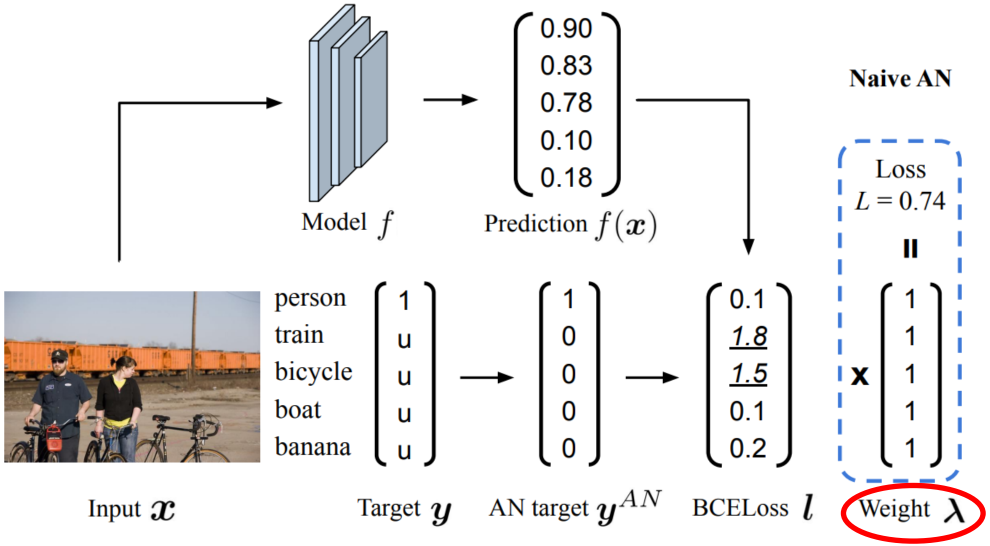
Our method



Define AN target $y_i^{AN} = \begin{cases} 1, & i \in \mathcal{S}^p \\ 0, & i \in \mathcal{S}^n \cup \mathcal{S}^u \end{cases}$ where

$$\mathcal{S}^p = \{i \mid y_i = 1\}$$
$$\mathcal{S}^n = \{i \mid y_i = 0\}$$
$$\mathcal{S}^u = \{i \mid y_i = u\}$$

Our method

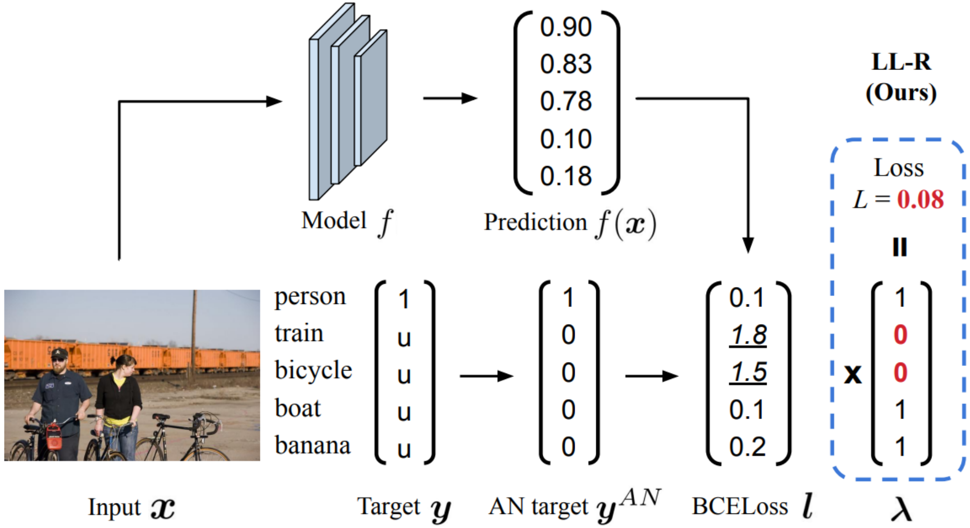


Introduce the weight term λ_i in a standard BCE loss function

$$L = \frac{1}{|\mathcal{D}'|} \sum_{(\boldsymbol{x}, \boldsymbol{y}^{AN}) \in \mathcal{D}'} \frac{1}{K} \sum_{i=1}^{K} \mathrm{BCELoss}\left(f(\boldsymbol{x})_i, y_i^{AN}\right) \times \lambda_i$$

Naïve AN (Vanilla BCE) : $\lambda_i = 1$ for all i

Our method

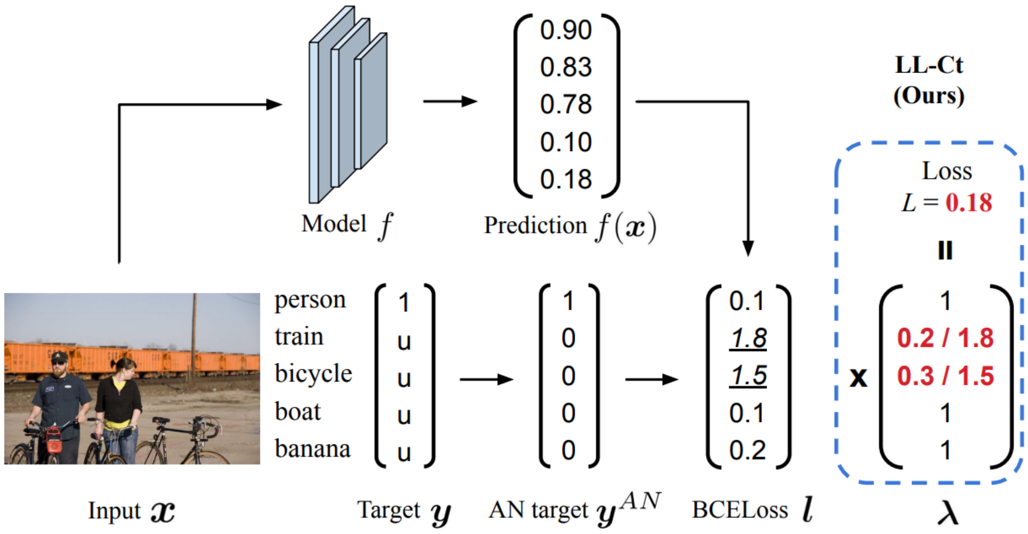


1) LargeLoss-Rejection (LL-R)

$$\lambda_i = \begin{cases} 0, & i \in \mathcal{S}^u \text{ and } l_i > R(t) \\ 1, & \text{otherwise} \end{cases}$$

$R(t)$: Top $[(t-1) \cdot \Delta_{rel}]\%$ loss value in mini-batch at epoch t

Our method

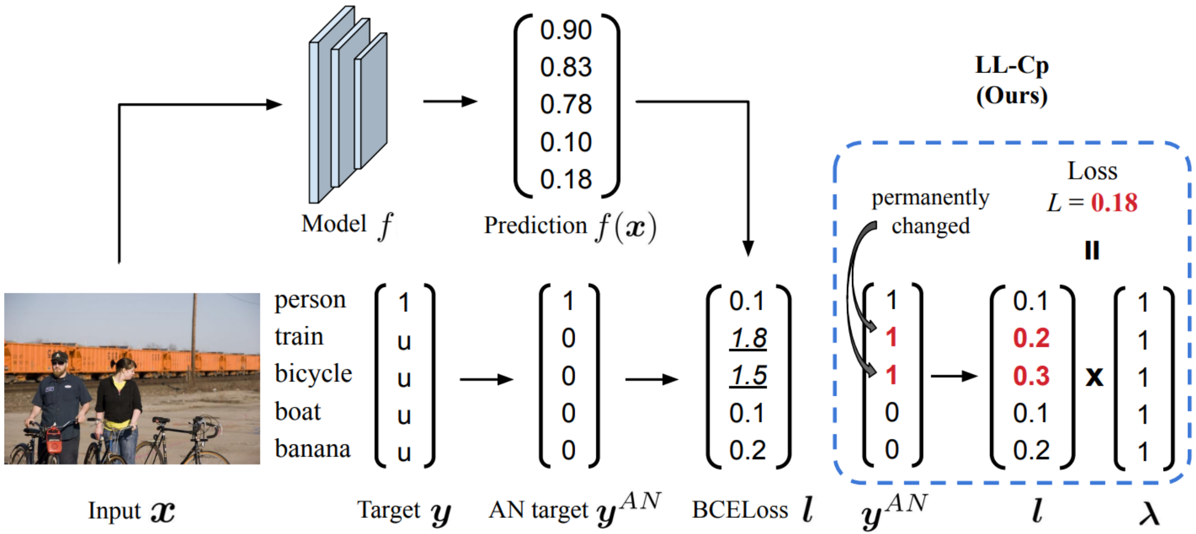


2) LargeLoss-Correction(temporary) (LL-Ct)

$$\lambda_i = \begin{cases} \frac{\log f(\boldsymbol{x})_i}{\log(1 - f(\boldsymbol{x})_i)}, & i \in \mathcal{S}^u \text{ and } l_i > R(t) \\ 1, & \text{otherwise} \end{cases}$$

$R(t)$: Top $[(t-1) \cdot \Delta_{rel}]\%$ loss value in mini-batch at epoch t

Our method



3) LargeLoss-Correction(permanent) (LL-Cp)

$$\lambda_i = 1 \text{ for all } i \quad \text{with} \quad y_i^{AN} = \begin{cases} 1, & i \in \mathcal{S}^u \text{ and } l_i > R(t) \\ unchanged, & \text{otherwise} \end{cases}$$

$R(t)$: Top $[\Delta_{rel}]\%$ loss value in mini-batch at epoch t

Results

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

1) In artificially created partial label datasets

Method	End-to-end				LinearInit.			
	VOC	COCO	NUSWIDE	CUB	VOC	COCO	NUSWIDE	CUB
Full label	90.2	78.0	54.5	32.9	91.1	77.2	54.9	34.0
Naive AN	85.1	64.1	42.0	19.1	86.9	68.7	47.6	20.9
WAN [7, 27]	86.5	64.8	46.3	20.3	87.1	68.0	47.5	21.1
LSAN [7, 37]	86.7	66.9	44.9	17.9	86.5	69.2	50.5	16.6
EPR [7]	85.5	63.3	46.0	20.0	84.9	66.8	48.1	21.2
ROLE [7]	87.9	66.3	43.1	15.0	88.2	69.0	**51.0**	16.8
LL-R (Ours)	**89.2**	**71.0**	47.4	19.5	**89.4**	**71.9**	49.1	21.5
LL-Ct (Ours)	89.0	70.5	48.0	**20.4**	89.3	71.6	49.6	**21.8**
LL-Cp (Ours)	88.4	70.7	**48.3**	20.1	88.3	71.0	49.4	21.4

Results

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

2) In a real partial label dataset (OpenImages V3)

Method	G1	G2	G3	G4	G5	All Gs
Naive IU	69.5	70.3	74.8	79.2	85.5	75.9
Curriculum [9]	70.4	71.3	76.2	80.5	86.8	77.1
IMCL [16]	71.0	72.6	77.6	81.8	87.3	78.1
Naive AN	77.1	78.7	81.5	84.1	88.8	82.0
WAN [7, 27]	71.8	72.8	76.3	79.7	84.7	77.0
LSAN [7, 37]	68.4	69.3	73.7	77.9	85.6	75.0
LL-R (Ours)	77.4	79.1	82.0	84.5	89.5	82.5
LL-Ct (Ours)	77.7	79.3	82.1	84.7	89.4	**82.6**
LL-Cp (Ours)	77.6	79.1	81.9	84.6	89.4	82.5

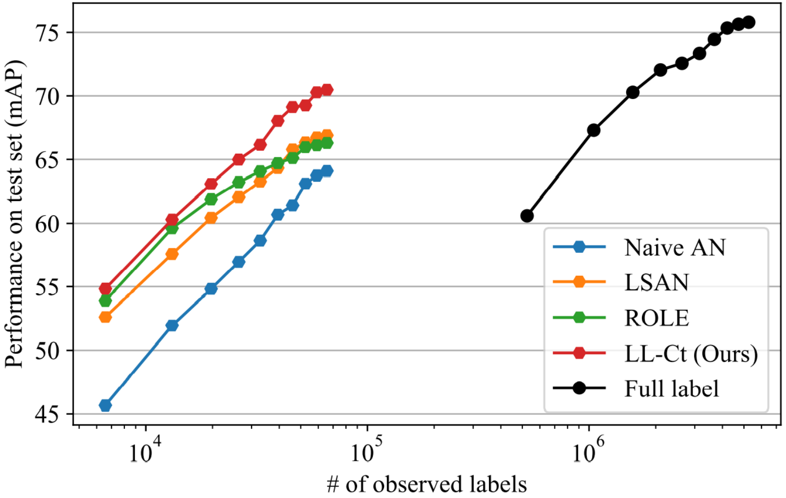
Qualitative results

Given : banana
➡ banana, orange
➡ banana, orange, bowl

GT : banana, orange, bowl

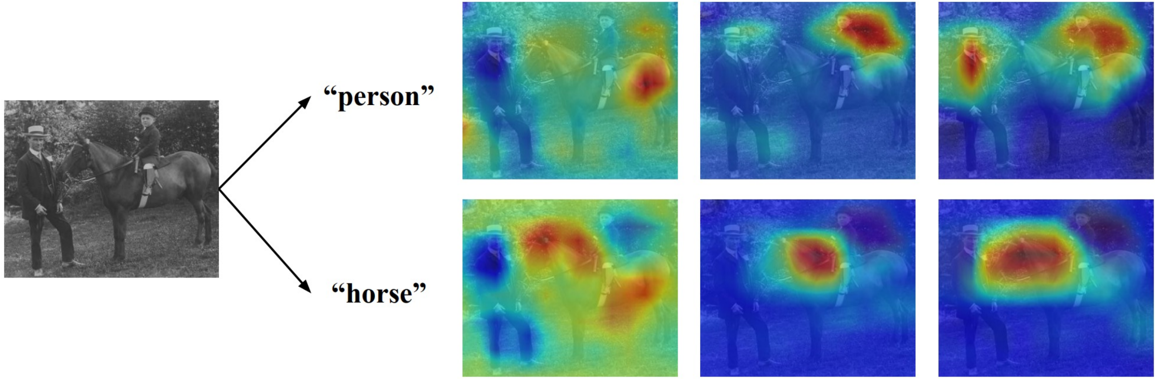
Given : vase
➡ vase, person
➡ vase, person, chair
➡ vase, person, chair, dining table

GT : vase, person, chair, dining table, bottle, wine glass



Analysis

CAM visualization



(a) Original image (b) GT category (c) Naive AN (d) ROLE (e) LL-Ct (Ours)

Pointing game result

Method	VOC	COCO
Naive AN	78.9	46.4
WAN [7, 28]	79.8	47.7
LSAN [7, 39]	79.5	49.1
EPR [7]	80.2	48.1
ROLE [7]	82.5	51.5
LL-R (Ours)	**83.7**	54.0
LL-Ct (Ours)	**83.7**	**54.1**
LL-Cp (Ours)	83.5	53.3

Conclusion

- In this paper, we present a **large loss modification scheme** that rejects or corrects the large loss samples appearing during training the multi-label classification model with partially labeled annotation.

- This originates from our empirical observation that **memorization effect** also happens in a noisy multi-label classification scenario.

- Although heavy and complex components are not included, our scheme successfully keeps the multi-label classification model from memorizing the noisy false negative labels, achieving **state-of-the-art performance** on various partially labeled multi-label datasets.

THANK YOU!

Code available!