南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Large Loss Matters in Weakly Supervised Multi-Label Classification

Youngwook Kim[1*] Jae Myung Kim[2*] Zeynep Akata[2,3,4] Jungwoo Lee[1,5]

[1]Seoul National University [2]University of Tübingen
[3]Max Planck Institute for Intelligent Systems [4]Max Planck Institute for Informatics [5]HodooAI Lab

CVPR 2022

Background

The arrow indicates the change of categories with positive label during training in our correction scheme LL-Ct and GT indicates actual ground truth positive labels for a training image. We show three cases where LL-Ct modifies the unannotated ground truth label correctly, and the failure case at the fourth column.

Given : fire hydrant
➡ fire hydrant, car
➡ fire hydrant, car, person, bicycle

GT : fire hydrant, car, person, bicycle

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

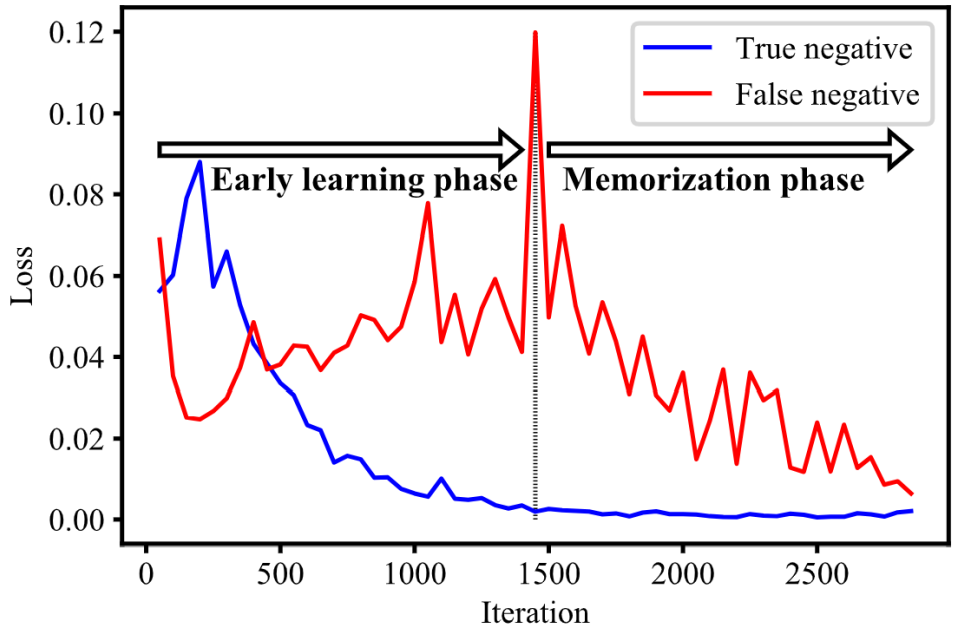
Given : truck
➡ truck, car
➡ truck, car, person

GT : truck, boat, motorcycle

We report the failure case of our method on the rightmost side where the model confuses the car as truck which is a similar category and mis-understands the absent category person as present

Background

Memorization in WSML. When training ResNet-50 model on PASCAL VOC dataset with partial label, we set all un-observed labels as negative. These labels are composed of true negative and false negative. We observe that the model first fits into true negative label (learning), and then fits into false negative (memorization)



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**

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).

Methods

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

We start the method with Assume Negative (AN) where all the unknown labels are regarded as negative. We call this modified target as:

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

The naive way of training the model with the dataset D is to minimize the loss function:

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

Large Loss Modification: $L = \frac{1}{|\mathcal{D}'|} \sum_{(\boldsymbol{x}, \boldsymbol{y}^{AN}) \in \mathcal{D}'} \frac{1}{K} \sum_{i=1}^{K} l_i \times \lambda_i$

$l_i = \text{BCELoss}\left(f(\boldsymbol{x})_i, y_i^{AN}\right)$

λ_i should be small when $i \epsilon S^u$ and the loss l_i has high value in the middle of the training, that is, to ignore that loss since it is likely to be the loss from a false negative sample. We set $\lambda_i = 1$ when $i \epsilon S^p \cup S^n$ since the label y_i^{AN} from these indices is a clean label i.

Methods

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Large loss rejection (LL-R):

$$\lambda_i = \begin{cases} 0, & i \in \mathcal{S}^u \text{ and } l_i > R(t) \\ 1, & \text{otherwise}, \end{cases} \qquad [(t-1)\cdot\Delta_{rel}]\%$$

Large loss correction (temporary) (LL-Ct):

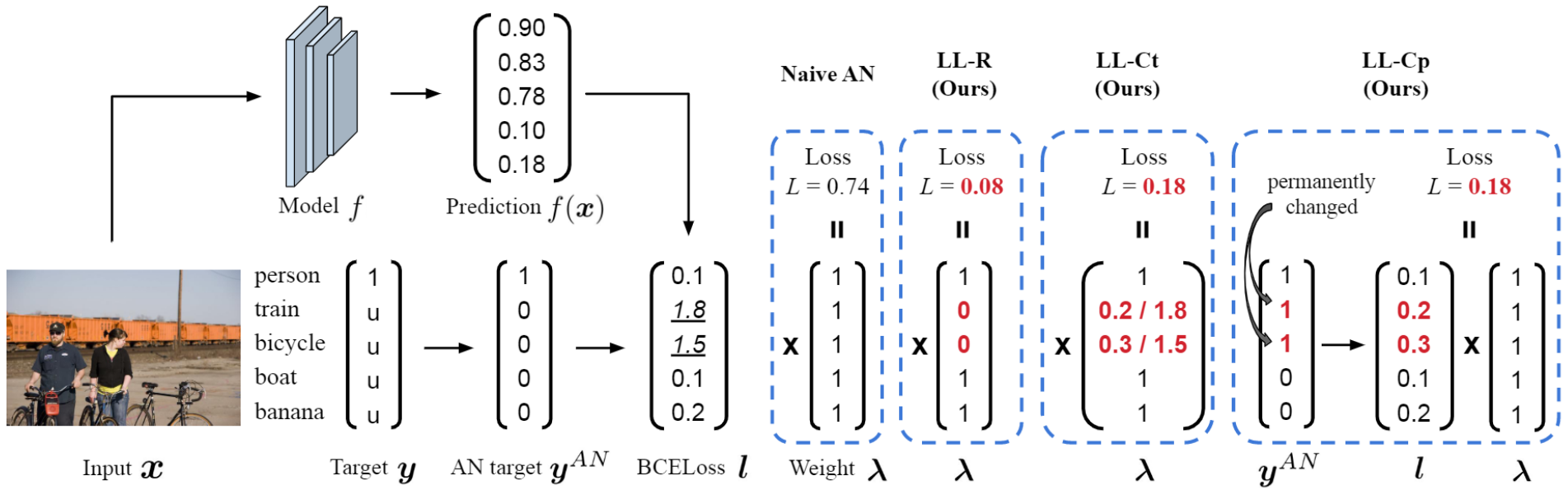
$$\lambda_i = \begin{cases} \frac{\log f(\boldsymbol{x})_i}{\log(1-f(\boldsymbol{x})_i)}, & i \in \mathcal{S}^u \text{ and } l_i > R(t) \\ 1, & \text{otherwise}, \end{cases} \qquad R(t) = R_0 - t\cdot\Delta_{abs}.\ R_0$$

$$\begin{aligned} l_i \times \lambda_i &= \text{BCELoss}(f(\boldsymbol{x})_i, y_i^{AN} = 0) \times \lambda_i \\ &= -\log(1-f(\boldsymbol{x})_i) \times \lambda_i \\ &= -\log f(\boldsymbol{x})_i \\ &= \text{BCELoss}(f(\boldsymbol{x})_i, 1)\,. \end{aligned}$$

Large loss correction (permanent) (LL-Cp):

$$y_i^{AN} = \begin{cases} 1, & i \in \mathcal{S}^u \text{ and } l_i > R(t) \\ \textit{unchanged}, & \text{otherwise}, \end{cases} \qquad R(t) = R_0 - t\cdot\Delta_{abs}.\ R_0$$

$$\mathcal{S}^u \leftarrow \mathcal{S}^u - \{i\}$$
$$\mathcal{S}^p \leftarrow \mathcal{S}^p \cup \{i\}$$



Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

"End-to-end" indicates that the entire weights of the model is fine-tuned from the beginning, while "LinearInit." indicates the backbone is frozen for the first few epochs.

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, 28]	86.5	64.8	46.3	20.3	87.1	68.0	47.5	21.1
LSAN [7, 39]	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

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, 28]	71.8	72.8	76.3	79.7	84.7	77.0
LSAN [7, 39]	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

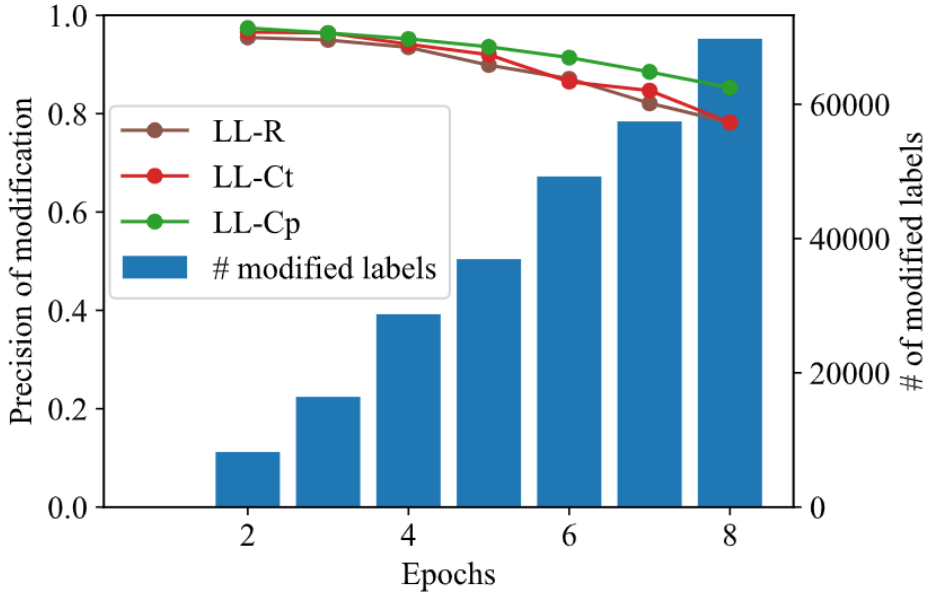
quantitative results in OpenImages V3 dataset with real partial label.

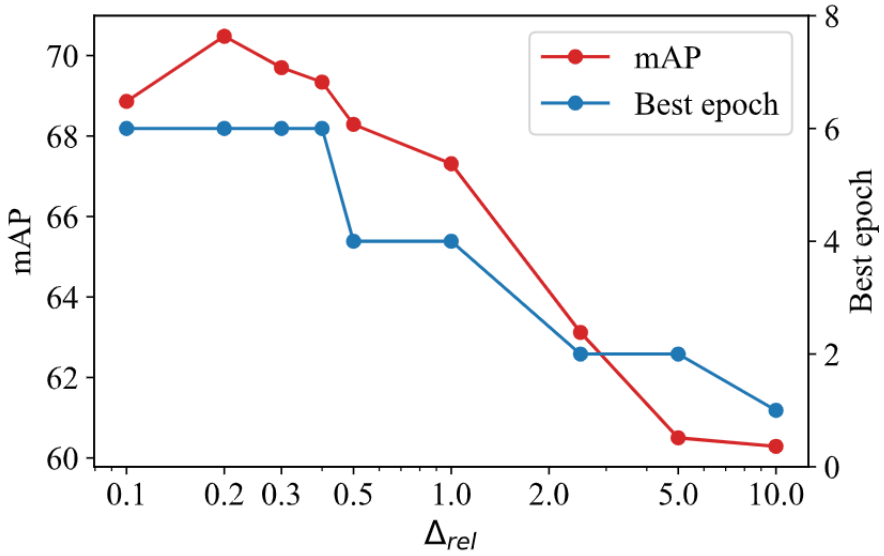
Cole E, Mac Aodha O, Lorieul T, et al. Multi-label learning from single positive labels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 933-942.

Experiments

Precision analysis of proposed methods on COCO dataset. Among the labels modified by our scheme as its loss values are large, we calculate the percentage of labels whose actual label is positive. We observe that our schemes indeed modify the false negative labels with high precision.

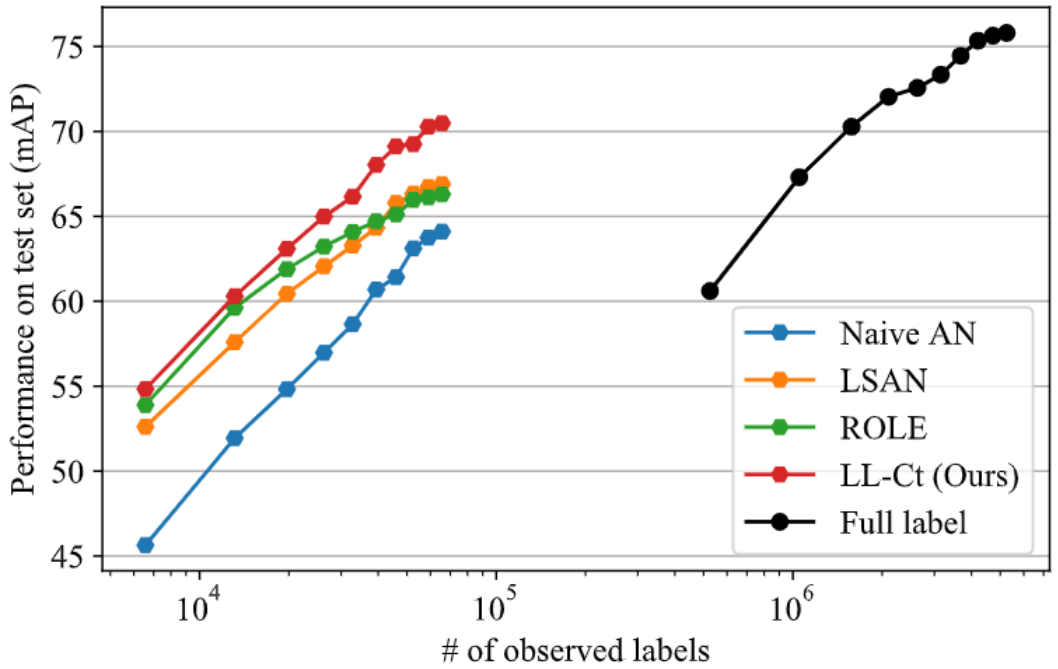
Hyperparameter effect of LL-Ct on COCO dataset. We observe that the model produces the best mAP when $\Delta rel = 0.2$.





Experiments

The number of observed labels for weakly supervised methods with 100% of training image is much more smaller than the fully supervised method with 10% of training image



Cole E, Mac Aodha O, Lorieul T, et al. Multi-label learning from single positive labels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 933-942.

Thanks