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

Multi-Label Learning from Single Positive Labels

Elijah Cole[1] Oisin Mac Aodha[2] Titouan Lorieul[3] Pietro Perona[1] Dan Morris[4] Nebojsa Jojic[5]
[1]Caltech [2]University of Edinburgh [3]Inria [4]Microsoft AI for Earth [5]Microsoft Research

CVPR 2021

Background

We extend existing multi-label losses to this setting and constrain the number of expected positive labels during training :

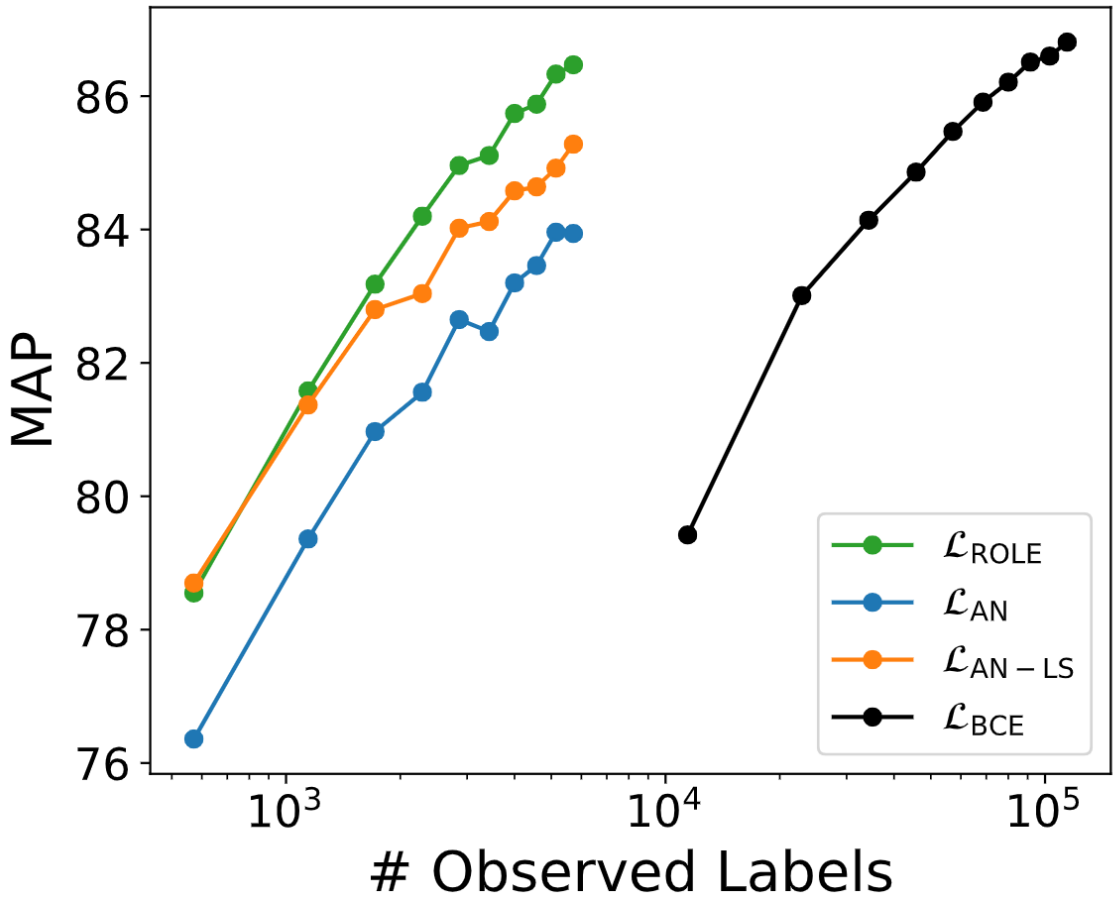
1. An effective method for this setting could allow for significantly reduced annotations costs for future datasets.

2. Multi-class datasets may have images that actually contain more than one class. For instance, the iNaturalist dataset has many images of insects on plants, but only one is annotated as the true class.

3. It is of scientific interest to understand how well multi-label classifiers can be made to perform at the minimal limit of supervision.

Background

Our experiments show that training with a single positive label per image allows us to drastically reduce the amount of supervision required to train multi-label image classifiers, while only incurring a tolerable drop in classification performance



L_{BCE} receives all 20 labels per image, while the other methods only receive one positive label per training image. Despite having a factor of 20 times fewer labels, our L_{ROLE} approach achieves comparable performance to the fully labeled case (L_{BCE})

Background

For Fully Observed Labels :

$$\mathcal{L}_{\mathrm{BCE}}(\mathbf{f}_n, \mathbf{y}_n) = -\frac{1}{L}\sum_{i=1}^{L}[\mathbb{1}_{[y_{ni}=1]}\log(f_{ni}) + \mathbb{1}_{[y_{ni}=0]}\log(1-f_{ni})]$$

$P(y_i = 1|\mathbf{x}_n)$

For Partially Observed Labels :

unobserved

$\mathbf{z}_n \in \mathcal{Z} = \{0, 1, \varnothing\}^L$

$$\mathcal{L}_{\mathrm{IU}}(\mathbf{f}_n, \mathbf{z}_n) = -\frac{1}{L}\sum_{i=1}^{L}[\mathbb{1}_{[z_{ni}=1]}\log(f_{ni}) + \mathbb{1}_{[z_{ni}=0]}\log(1-f_{ni})]$$

ignores unobserved negatives (IUN)

$$\mathcal{L}_{\mathrm{IUN}}(\mathbf{f}_n, \mathbf{z}_n, \mathbf{y}_n) = -\frac{1}{L}\sum_{i=1}^{L}[\mathbb{1}_{[z_{ni}=1]}\log(f_{ni}) + \mathbb{1}_{[y_{ni}=0]}\log(1-f_{ni})]$$

Positive Only Labels :

$z_{ni} \in \{1, \varnothing\}$ for all $n \in \{1, \ldots, N\}, i \in \{1, \ldots, L\}$

$\sum_{i=1}^{L} \mathbb{1}_{[z_{ni}=1]} = 1$ for all $n \in \{1, \ldots, N\}$

$$\mathcal{L}_{\mathrm{AN}}(\mathbf{f}_n, \mathbf{z}_n) = -\frac{1}{L}\sum_{i=1}^{L}[\mathbb{1}_{[z_{ni}=1]}\log(f_{ni}) + \mathbb{1}_{[z_{ni}\neq 1]}\log(1-f_{ni})]$$

Methods

Sum of labels: L

Weak negatives：

$$\mathcal{L}_{\mathrm{WAN}}(\mathbf{f}_n, \mathbf{z}_n) = -\frac{1}{L}\sum_{i=1}^{L}[\mathbb{1}_{[z_{ni}=1]}\log(f_{ni}) + \mathbb{1}_{[z_{ni}\neq 1]}\gamma\log(1-f_{ni})]$$

$\gamma = 1/(L-1)$

pseudo-negative sampling:

$$-\frac{1}{L}\sum_{i=1}^{L}[\mathbb{1}_{[z_{ni}=1]}\log(f_{ni}) + \mathbb{1}_{[z_{ni}\neq 1]}\eta_{ni}\log(1-f_{ni})]$$

Label Smoothing：

Def: replaces $\mathbf{y}_n$ with $(1-\epsilon)\mathbf{y}_n + \epsilon\mathbf{u}$

$[1/L, \ldots, 1/L]$

$$\mathcal{L}_{\mathrm{AN-LS}}(\mathbf{f}_n, \mathbf{z}_n) = -\frac{1}{L}\sum_{i=1}^{L}[\mathbb{1}^{\frac{\epsilon}{2}}_{[z_{ni}=1]}\log(f_{ni}) + \mathbb{1}^{\frac{\epsilon}{2}}_{[z_{ni}\neq 1]}\log(1-f_{ni})]$$

Methods

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

Positive Regularization :

$$k = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\text{data}}(\mathbf{x},\mathbf{y})} \sum_{i=1}^{L} \mathbb{1}_{[y_i=1]} \qquad \hat{k}(\mathbf{F}_B) = \frac{\sum_{n\in B}\sum_{i=1}^{L} \mathbf{f}_{ni}}{|B|}$$

courages $\hat{k}(\mathbf{F}_B)$ to be close to k

$$\mathcal{L}_{\text{EPR}}(\mathbf{F}_B, \mathbf{Z}_B) = \frac{1}{|B|}\sum_{n\in B} \mathcal{L}^{+}_{\text{BCE}}(\mathbf{f}_n, \mathbf{z}_n) + \lambda R_k(\mathbf{F}_B)$$

$$R_k(\mathbf{F}_B) = \left(\frac{\hat{k}(\mathbf{F}_B) - k}{L}\right)^2$$, the squared error

Online Estimation of Unobserved Labels :

jointly train the label estimator $g(\cdot;\phi)$ and the image classifier $f(\cdot;\theta)$

We write the estimated labels as : $\tilde{\mathbf{Y}} \in [0,1]^{N\times L}$

$$\mathcal{L}'(\mathbf{F}_B|\tilde{\mathbf{Y}}_B) = \frac{1}{|B|}\sum_{n\in B}\mathcal{L}_{\mathrm{BCE}}(\mathbf{f}_n, \mathrm{sg}(\tilde{\mathbf{y}}_n)) + \mathcal{L}_{\mathrm{EPR}}(\mathbf{F}_B, \mathbf{Z}_B)$$

the stop-gradient function

the label estimator $g(\mathbf{x}_n;\phi)$

The L_{BCE} term encourages the image classifier predictions F_B to match the estimated labels $\tilde{Y}_B$, while the L_{EPR} term pushes F_B to correctly predict known positives and respect the expected number of positives per image.

update θ while assuming that ϕ is fixed

update ϕ while assuming θ is fixed

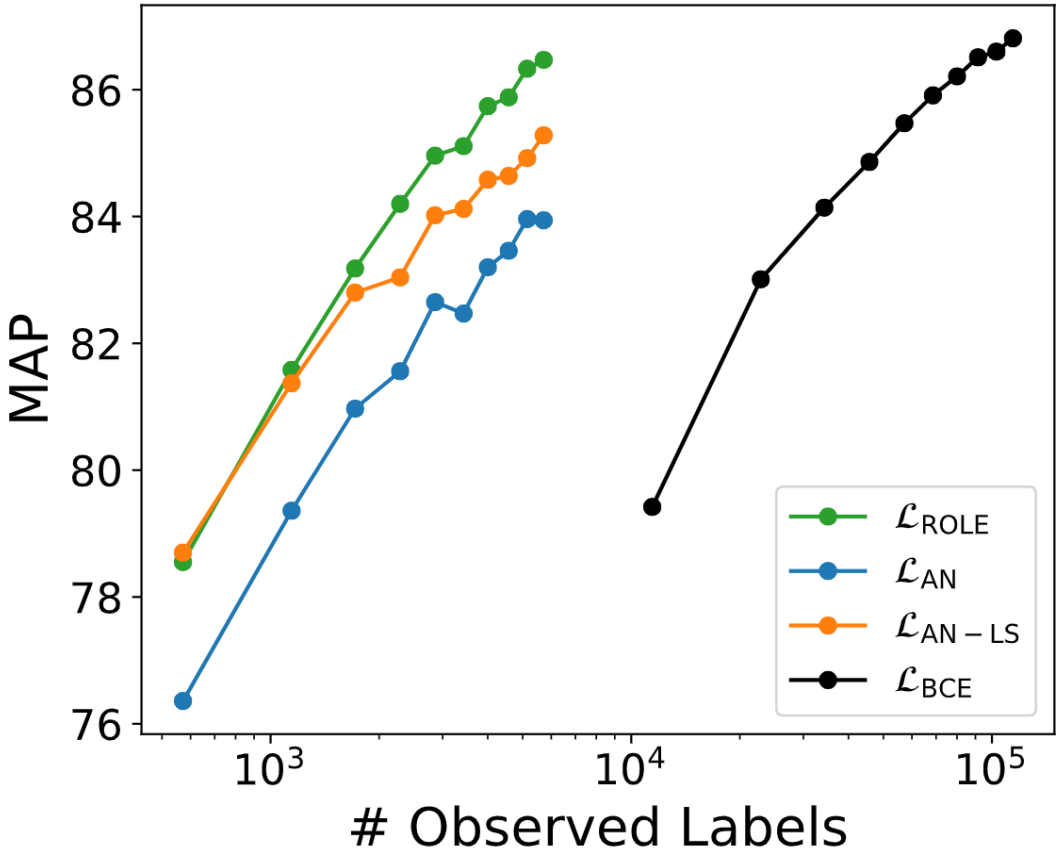
$$\mathcal{L}_{\mathrm{ROLE}}(\mathbf{F}_B, \tilde{\mathbf{Y}}_B) = \frac{\mathcal{L}'(\mathbf{F}_B|\tilde{\mathbf{Y}}_B) + \mathcal{L}'(\tilde{\mathbf{Y}}_B|\mathbf{F}_B)}{2}$$

Experiments

Multi-label test set mean average precision (MAP) for different multi-label losses on four different image classification datasets.
We present results for two scenarios: (i) training a linear classifier on fixed features and (ii) fine-tuning the entire network end-to-end.
For losses labeled with "LinearInit." we freeze the weights of the backbone network for the initial epochs of training and then fine-tune the entire network end-to-end for the remaining epochs.

		Linear				Fine-Tuned			
Loss	Labels Per Image	VOC12	COCO	NUS	CUB	VOC12	COCO	NUS	CUB
$\mathcal{L}_{\text{BCE}}$	All Pos. & All Neg.	86.7	70.0	50.7	29.1	89.1	75.8	52.6	32.1
$\mathcal{L}_{\text{BCE-LS}}$	All Pos. & All Neg.	87.6	70.2	51.7	29.3	90.0	76.8	53.5	32.6
$\mathcal{L}_{\text{IUN}}$	1 Pos. & All Neg.	86.4	67.0	49.0	19.4	87.1	70.5	46.9	21.3
$\mathcal{L}_{\text{IU}}$	1 Pos. & 1 Neg.	82.6	60.8	43.6	16.1	83.2	59.7	42.9	17.9
$\mathcal{L}_{\text{AN}}$	1 Pos. & 0 Neg.	84.2	62.3	46.2	17.2	85.1	64.1	42.0	19.1
$\mathcal{L}_{\text{AN-LS}}$	1 Pos. & 0 Neg.	<u>85.3</u>	<u>64.8</u>	<u>48.5</u>	15.4	86.7	66.9	44.9	17.9
$\mathcal{L}_{\text{WAN}}$	1 Pos. & 0 Neg.	84.1	63.1	45.8	<u>17.9</u>	86.5	64.8	46.3	**20.3**
$\mathcal{L}_{\text{EPR}}$	1 Pos. & 0 Neg.	83.8	62.6	46.4	**18.0**	85.5	63.3	46.0	<u>20.0</u>
$\mathcal{L}_{\text{ROLE}}$	1 Pos. & 0 Neg.	**86.5**	**66.3**	**49.5**	16.2	<u>87.9</u>	66.3	43.1	15.0
$\mathcal{L}_{\text{AN-LS}}$ +LinearInit.	1 Pos. & 0 Neg.	-	-	-	-	86.5	**69.2**	<u>50.5</u>	16.6
$\mathcal{L}_{\text{ROLE}}$ +LinearInit.	1 Pos. & 0 Neg.	-	-	-	-	**88.2**	<u>69.0</u>	**51.0**	16.8

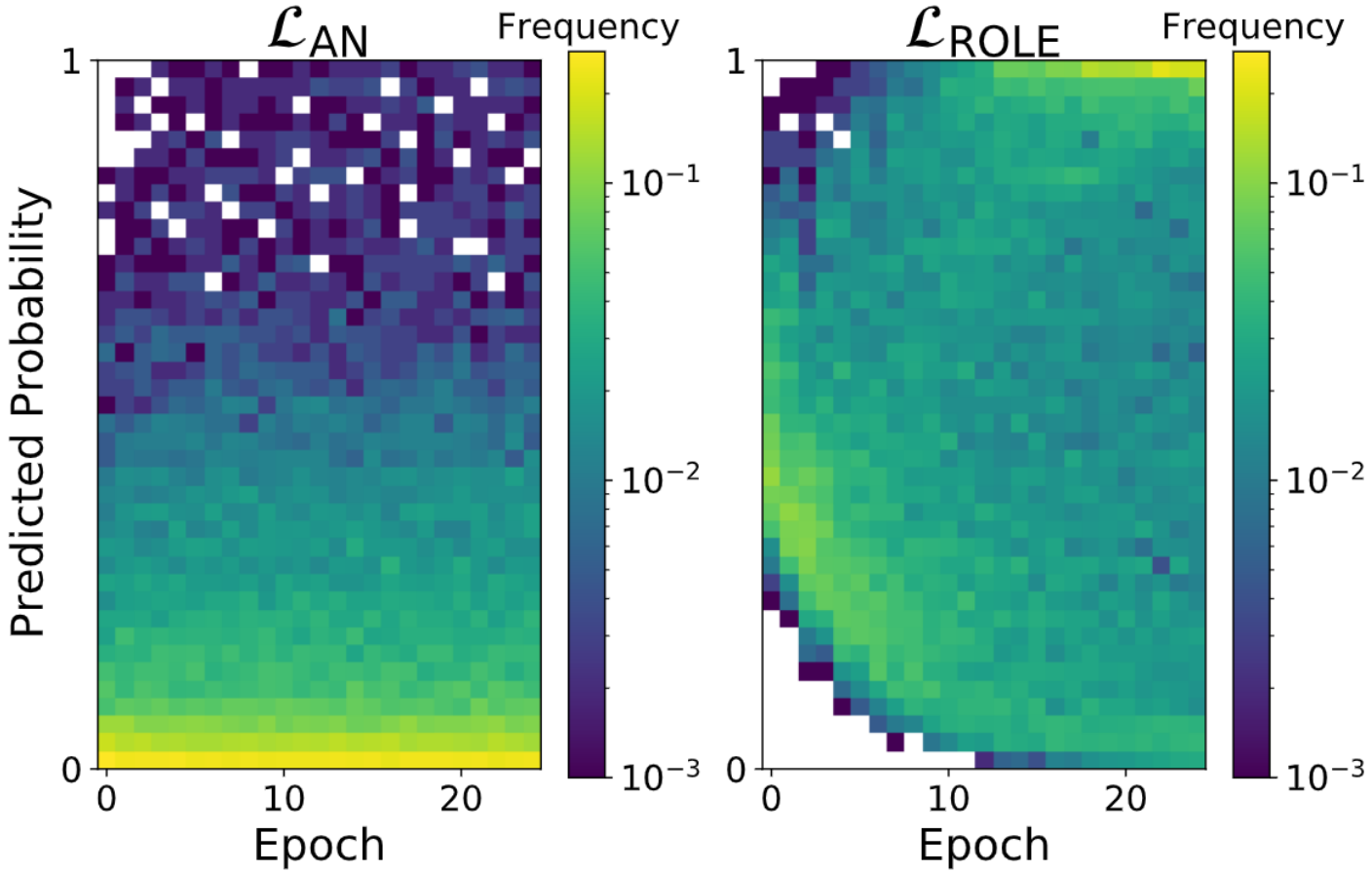
Experiments



Loss	VOC12	COCO	NUS	CUB
$\mathcal{L}_{\text{AN}}$	85.8	63.8	49.3	16.8
$\mathcal{L}_{\text{AN}-\text{LS}}$	86.9	65.4	49.7	17.4
$\mathcal{L}_{\text{ROLE}}$	**90.3**	**69.5**	**56.0**	**19.6**

Training set MAP for multi-label predictions evaluated with respect to the full ground truth labels. These values measure how well each method recovers the true training labels despite being trained with one positive label per image.

Experiments



Distribution of predicted probabilities for unobserved positives when training with a single positive per image for COCO. Each column represents a normalized histogram and white pixels indicate a frequency of zero. Training with L_{ROLE} (right) results in the recovery of a significant number of the unlabeled positives as evident by the majority of the probability correctly being con- centrated at 1.0 (top right) by the end of training. L_{AN} (left) does not exhibit the same behavior.

Thanks