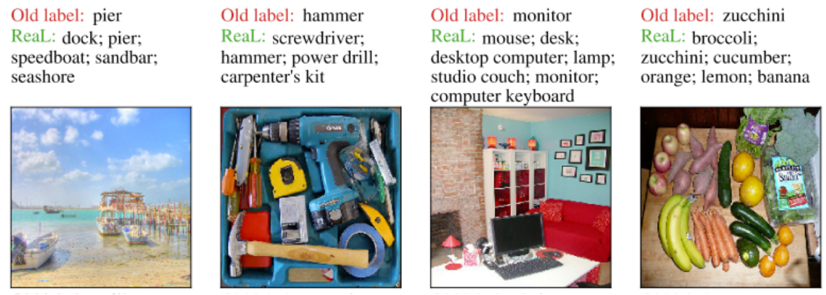
Tackling the biases in Single-Positive Multi-Label Learning

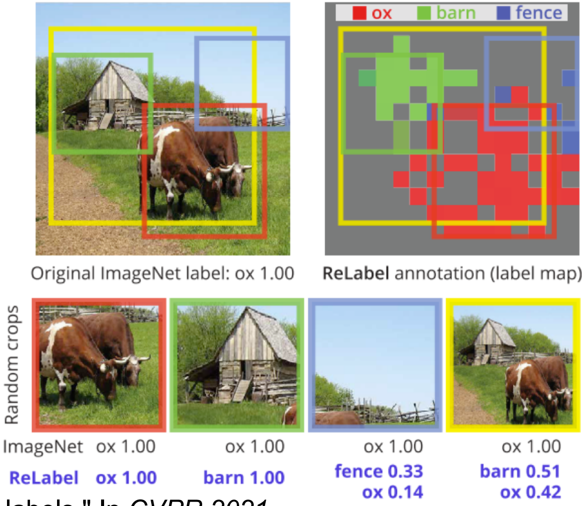
Meng-Jiun Chiou[1,2], **Henghui Ding**[2], Hanshu Yan[1],
Junhao Liew[2] and Roger Zimmermann[1]

[1]National University of Singapore [2]**ByteDance AI Lab**

Why Single-Positive Multi-Label (SPML)?

- Existing multi-class classification dataset neglects the fact of multi-label
- Some works thus propose to exhaustively annotate (at least) validation sets[1]; other works propose to pseudo-label train sets[2]
- We focus on Single-Positive Multi-label (SPML)[3]: **only one positive label** is provided for each example





[1]Beyer, Lucas, et al. "Are we done with imagenet?." In *arXiv preprint arXiv:2006.07159* (2020).
[2]Yun, Sangdoo, et al. "Re-labeling imagenet: from single to multi-labels, from global to localized labels." In *CVPR 2021.*
[3]Cole, Elijah, et al. "Multi-Label Learning from Single Positive Labels." In *CVPR 2021.*

Recent Progress in SPML

- Generate pseudo-single-label datasets (randomly choose one label)

Recent Progress in SPML

- Generate pseudo-single-label datasets (randomly choose one label)
- Under the SPML setting, Cole et al.[1] propose
  - L_{EPR}: avoid the label noise by using BCE loss ONLY with the observed positive labels and regularizing the expected number of positive labels per image.

$$\mathcal{L}_{\text{EPR}}(\mathbf{F}_B, \mathbf{Z}_B) = \frac{1}{|B|} \sum_{n \in B} \mathcal{L}^{+}_{\text{BCE}}(\mathbf{f}_n, \mathbf{z}_n) + \lambda R_k(\mathbf{F}_B), \quad R_k(\mathbf{F}_B) = \left(\frac{\hat{k}(\mathbf{F}_B) - k}{L} \right)^2 \quad \hat{k}(\mathbf{F}_B) = \frac{\sum_{n \in B} \sum_{i=1}^{L} \mathbf{f}_{ni}}{|B|}.$$

  - L_{ROLE}: do **R**egularized **O**nline unobserved **L**abels **E**stimation

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

$\tilde{\mathbf{y}}_n = g(\mathbf{x}_n; \phi)$ where the *label estimator* $g: \mathcal{X} \to [0,1]^L$

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

An EM-like alternative training strategy
ref. "Bootstrap Your Own Latent (BYOL)"
by DeepMind at NeurIPS 2020

[1]Cole, Elijah, et al. "Multi-Label Learning from Single Positive Labels." In *CVPR 2021*.

Recent Progress in SPML - Results

- Train on fully-labeled/pseudo-SP set; test on fully-labeled set

mean Average Precision (mAP)	Pascal VOC (20 cls; 1.5 pos)		COCO (80 cls; 2.9 pos)		NUS-WIDE (81 cls; 1.9 pos)		CUB (312 cls; 31.4 pos)	
L_{BCE} (full labels)	86.7		70.0		51.6		29.0	
L_{AN}	83.8	-0.5	62.3	-3.7	45.9	-2.3	17.3	-12.6
L_{ROLE}	86.2		66.3		49.3		16.4	

The Flawful Single-Positive Labels

- The single-positive labels generation doesn't consider **labeling bias**: the larger objects/located at the center/more apparent (*bounded rationality*[1]), or easier to describe (*reporting bias*[2]), might be labeled more frequently.

[1]. H. A. Simon. "Bounded rationality." In *Utility and probability*. Springer, 1990.
[2]. I. Misra, et al. "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels." In *CVPR*, 2016.

The Flawful Single-Positive Labels

- The single-positive labels generation doesn't consider **labeling bias**: the larger objects/located at the center/more apparent (*bounded rationality*[1]), or easier to describe (*reporting bias*[2]), might be labeled more frequently.
- A naïve 3-step approach to introduce biases:
  1. Inference a model trained on fully-labeled datasets to generate top predictions
  2. Ranked predictions can be viewed as model's preference over classes
  3. The top-ranked ground truth label is selected as the single label of each image

[1] H. A. Simon. "Bounded rationality." In *Utility and probability*. Springer, 1990.
[2] I. Misra, et al. "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels." In *CVPR*, 2016.

Results of biased SPML

mean Average Precision (mAP)	Pascal VOC (20 cls; 1.5 pos)		COCO (80 cls; 2.9 pos)		NUS-WIDE (81 cls; 1.9 pos)		CUB (312 cls; 31.4 pos)	
L_{BCE} (full labels)	86.7		70.0		51.6		29.0	
L_{AN}	83.8	-0.5	62.3	-3.7	45.9	-2.3	17.3	-12.6
L_{ROLE}	86.2		66.3		49.3		16.4	
Biased L_{AN}	81.3	-2.0	48.2	-10.8	36.2	-8.5	15.7	-1.1
Biased L_{ROLE}	84.2		55.5		40.8		15.5	

[1]Chiou et al. “Recovering the Unbiased Scene Graphs from the Biased Ones.” In ACMMM 2021.
[2]Elkan et al. “Learning classifiers from only positive and unlabeled data”. In SIGKDD 2008.

Moving Forward: Toward unbiased SPML

- Modeling and removing the biases
  - **Positive-Unlabeled (PU) learning methods**
    - However, note that post-processing methods (e.g., DLFE[1]) don’t work for class-wise dataset-level metrics (e.g., mAP), since they do not change the ranking of per-class scores over a dataset

Preliminary Results

Evaluation metrics can themselves be the issues! Alternatives like (approx.) F1-score:

$$\frac{pr}{\Pr(\mathbf{y}=1)} = \frac{pr^2}{r\Pr(\mathbf{y}=1)} = \frac{\Pr(\mathbf{y}=1|\hat{\mathbf{y}}=1)r^2}{\Pr(\hat{\mathbf{y}}=1,\mathbf{y}=1)} = \frac{r^2}{\Pr(\hat{\mathbf{y}}=1)}.$$

Precision@3	Pascal VOC (20 cls; 1.5 pos) | biased test set		COCO (80 cls; 2.9 pos) | biased test set		NUS-WIDE (81 cls; 1.9 pos) | biased test		CUB (312 cls; 31.4 pos) | biased test set	
L_{BCE}	44.0		61.1		42.5		79.4	
L_{AN}	43.3		58.9		41.5		46.9	
Biased L_{AN}	42.3	32.6	52.1	31.9	41.0	31.2	76.8	28.3
Biased L_{AN} PU	42.7	32.6	52.2	31.8	40.2	30.8	52.6	18.4
L_{ROLE}	43.7		60.8		42.4		52.7	
Biased L_{ROLE}	42.7	32.8	55.5	32.4	41.3	31.1	76.5	27.9
Biased L_{ROLE} PU	43.0	32.8	50.9	31.1	39.2	30.1	60.9	11.1

Recall@3	Pascal VOC (20 cls; 1.5 pos) | biased test set		COCO (80 cls; 2.9 pos) | biased test set		NUS-WIDE (81 cls; 1.9 pos) | biased test		CUB (312 cls; 31.4 pos) | biased test set	
L_{BCE}	94.5		73.8		59.5		8.0	
L_{AN}	93.6		72.1		58.4		4.6	
Biased L_{AN}	92.3	97.7	64.8	95.6	57.6	93.5	7.7	84.9
Biased L_{AN} PU	92.8	97.9	64.8	95.4	56.7	92.2	5.3	55.1
L_{ROLE}	94.3		74.1		59.5		5.2	
Biased L_{ROLE}	93.0	98.5	69.0	97.3	58.2	93.3	7.7	83.8
Biased L_{ROLE} PU	93.3	98.4	63.5	93.3	55.6	90.3	6.2	33.3

Top@3	Pascal VOC (20 cls; 1.5 pos) | biased test set		COCO (80 cls; 2.9 pos) | biased test set		NUS-WIDE (81 cls; 1.9 pos) | biased test		CUB (312 cls; 31.4 pos) | biased test set	
L_{BCE}	98.2		97.6		73.3		95.5	
L_{AN}	98.1		97.4		73.0		81.4	
Biased L_{AN}	98.0	97.7	97.2	95.6	73.0	93.5	96.0	84.9
Biased L_{AN} PU	98.1	97.9	97.1	95.4	72.4	92.2	78.8	55.1
L_{ROLE}	98.5		98.0		73.5		85.7	
Biased L_{ROLE}	98.6	98.5	98.3	97.3	73.3	93.3	96.1	83.8
Biased L_{ROLE} PU	98.5	98.4	96.5	93.3	72.2	90.3	90.4	33.3

[1]Chiou et al. “Recovering the Unbiased Scene Graphs from the Biased Ones.” In ACMMM 2021.
[2]Elkan et al. “Learning classifiers from only positive and unlabeled data”. In SIGKDD 2008.

Moving Forward: Toward unbiased SPML

- Modeling and removing the biases
  - Positive-Unlabeled (PU) learning methods
    - However, note that post-processing methods (e.g., DLFE[1]) don’t work for class-wise dataset-level metrics (e.g., mAP), since they do not change the ranking of per-class scores over a dataset
    - **Need to choose other PU methods** like two-step/biased learning methods.[2]
  - **Model the more exact biases**, e.g., compute a class-wise size/relative location values to be offset

[1]Chiou et al. "Recovering the Unbiased Scene Graphs from the Biased Ones." In ACMMM 2021.
[2]Elkan et al. "Learning classifiers from only positive and unlabeled data". In SIGKDD 2008.

Moving Forward: Toward unbiased SPML

- **Introduce specific biases, instead of "black-box" biases**
  - Trained classifier can introduce biases; however, what kind of the biases remain unspecified.
  - More specific biases according to size/relative location of objects
- Modeling and removing the biases
  - Positive-Unlabeled (PU) learning methods
    - However, note that post-processing methods (e.g., DLFE[1]) don't work for class-wise dataset-level metrics (e.g., mAP), since they do not change the ranking of per-class scores over a dataset
    - Need to choose other PU methods like two-step/biased learning methods.[2]
  - Model the more exact biases, e.g., compute a class-wise size/relative location values to be offset

[1]Chiou et al. “Recovering the Unbiased Scene Graphs from the Biased Ones.” In ACMMM 2021.
[2]Elkan et al. “Learning classifiers from only positive and unlabeled data”. In SIGKDD 2008.

Moving Forward: Toward unbiased SPML

- Introduce specific biases, instead of “black-box” biases
  - Trained classifier can introduce biases; however, what kind of the biases remain unspecified.
  - More specific biases according to size/relative location of objects
- Modeling and removing the biases
  - Positive-Unlabeled (PU) learning methods
    - However, note that post-processing methods (e.g., DLFE[1]) don’t work for class-wise dataset-level metrics (e.g., mAP), since they do not change the ranking of per-class scores over a dataset
    - Need to choose other PU methods like two-step/biased learning methods.[2]
  - Model the more exact biases, e.g., compute a class-wise size/relative location values to be offset
- **Generalize beyond image classification**, e.g., video, texts