Open-World Semi-Supervised Learning

Kaidi Cao , Maria Brbic , Jure Leskovec

Published in ICLR 2022
79 citations as of today

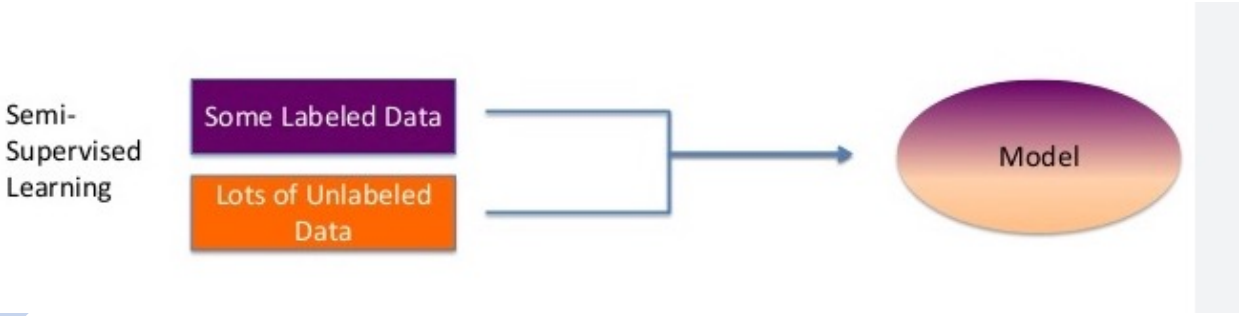
Presenter: Abdullah Mamun

Date: September 27, 2023

Motivation

- Suppose, you have a number of unique labels for a dataset.
- But the amount of labeled data in the dataset is very low. (e.g. <5%)

- We can use Semi-supervised learning to label the unlabeled data.
- But what about the unseen labels? E.g. eating in this case.



Open-World SSL

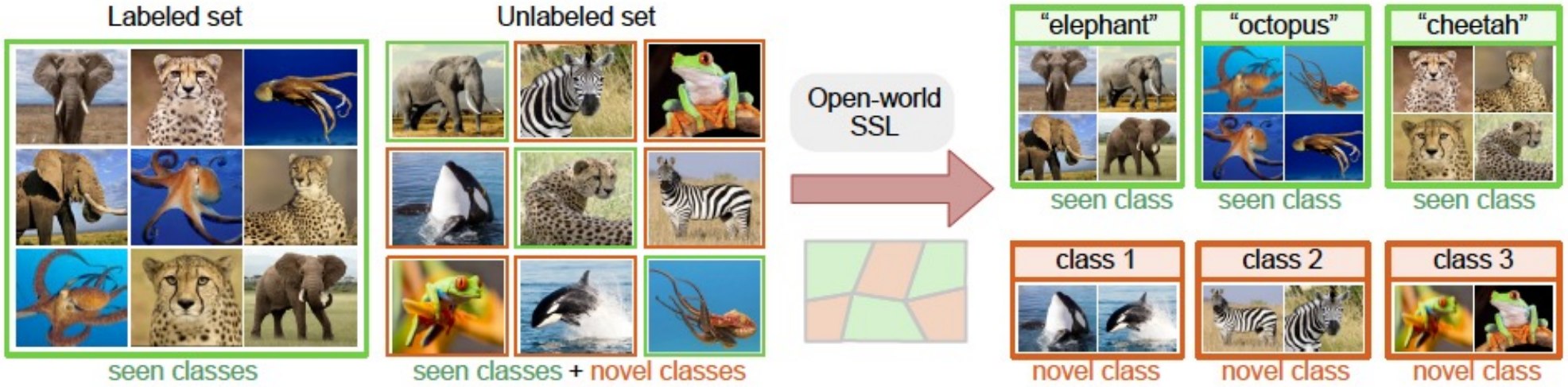


Figure 1: In the open-world SSL, the unlabeled dataset may contain classes that have never been encountered in the labeled set. Given unlabeled test set, the model needs to either assign instances to one of the classes previously seen in the labeled set, or form a novel class and assign instances to it.

Related Work: Novel class discovery

- The task is to cluster unlabeled dataset consisting of similar, but completely disjoint, classes than those present in the labeled dataset which is utilized to learn better representation for clustering.

- These methods assume that at the test time all the classes are novel.

- While these methods are able to discover novel classes, they do not recognize the seen/known classes. For the following figure, it will consider elephant, octopus, and cheetah as novel classes too.

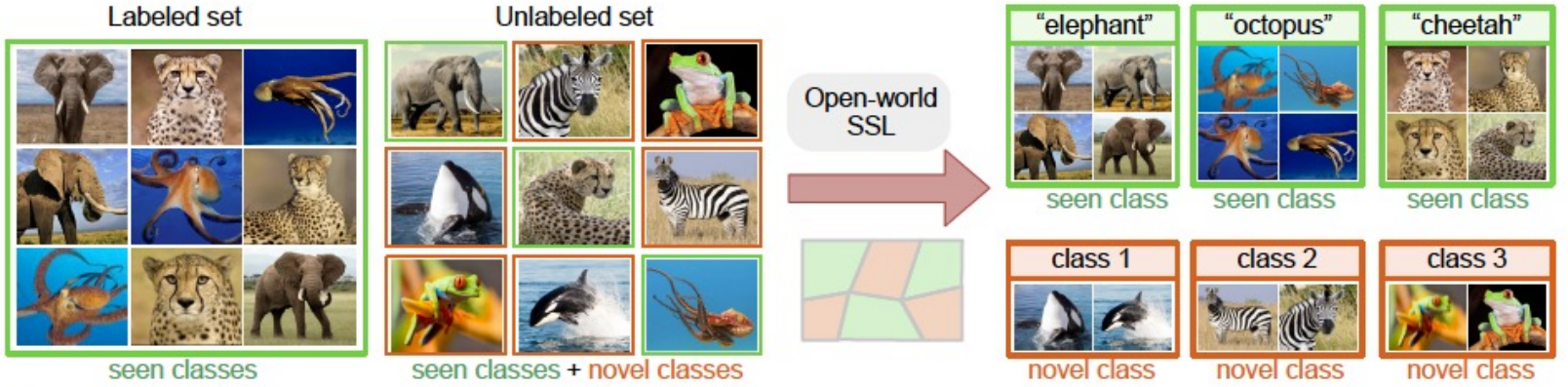
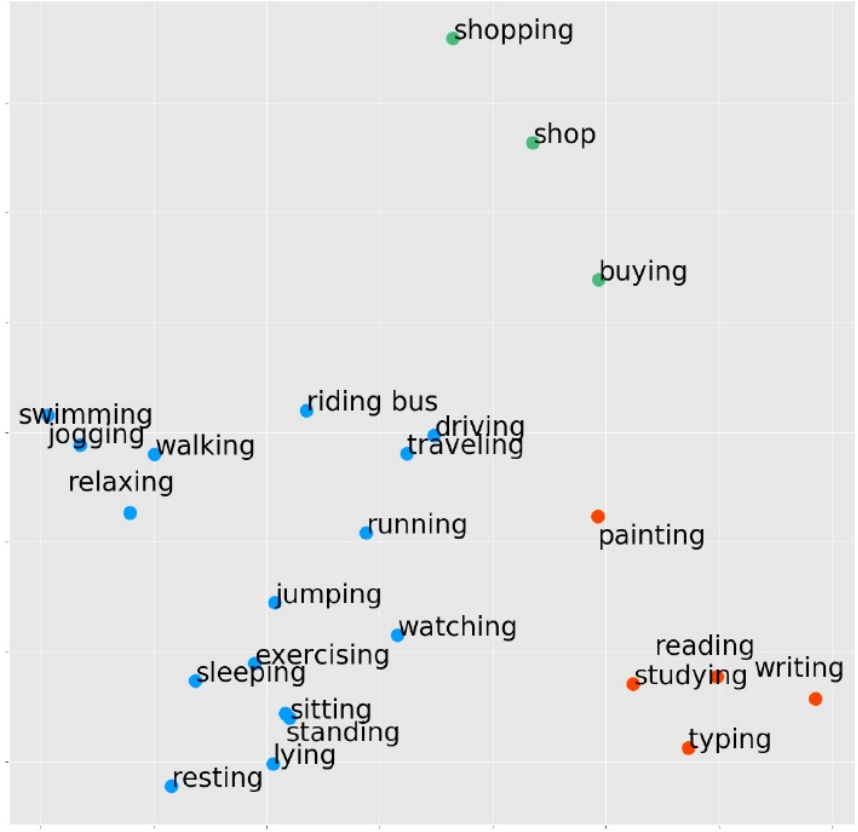


Figure 1: In the open-world SSL, the unlabeled dataset may contain classes that have never been encountered in the labeled set. Given unlabeled test set, the model needs to either assign instances to one of the classes previously seen in the labeled set, or form a novel class and assign instances to it.

Related Work: Traditional Semi-supervised learning

- Assumes closed-world setting in which labeled and unlabeled data come from the same set of classes.
- It will assume the unlabeled data can only belong to the classes seen in the labeled data.
- So, eating cannot be discovered in this method

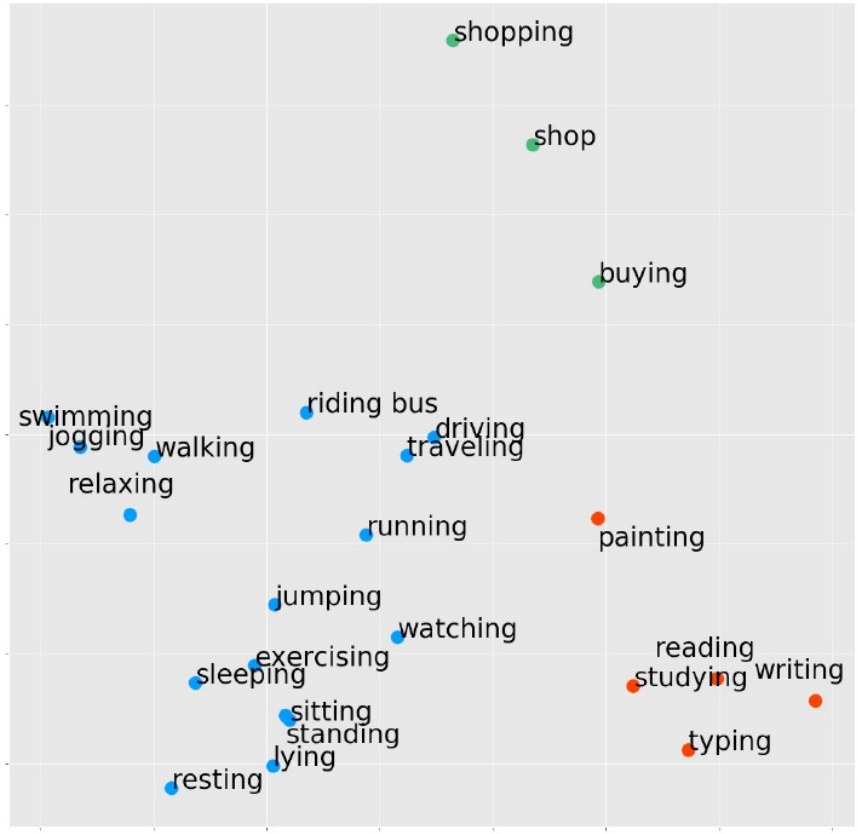


Related Work: Robust SSL

Robust SSL methods relax the SSL assumption by assuming that instances from novel classes may appear in the unlabeled test set.

The goal in robust SSL is to reject instances from novel classes which are treated as out-of-distribution instances.

- So, eating or any activities out of distribution will be rejected



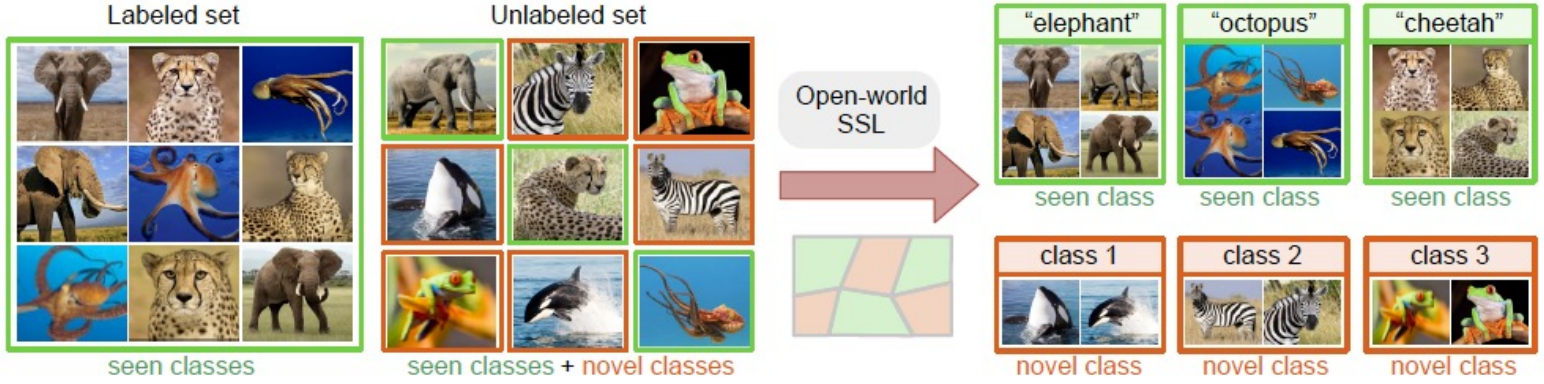
Related Work: Robust SSL

Open-world SSL can

- *Classify the seen classes,*
- *Discover novel classes, &*
- *Requires no prior knowledge*

Table 1: Relationship between our novel open-world SSL and other machine learning settings.

Setting	Seen classes	Novel classes	Prior knowledge
Novel class discovery	Not present	Discover	None
SSL	Classify	Not present	None
Robust SSL	Classify	Reject	None
Generalized zero-shot learning	Classify	Discover	Class attributes
Open-set recognition	Classify	Reject	None
Open-world recognition	Classify	Discover	Human-in-the-loop
Open-world SSL	Classify	Discover	None



ORCA

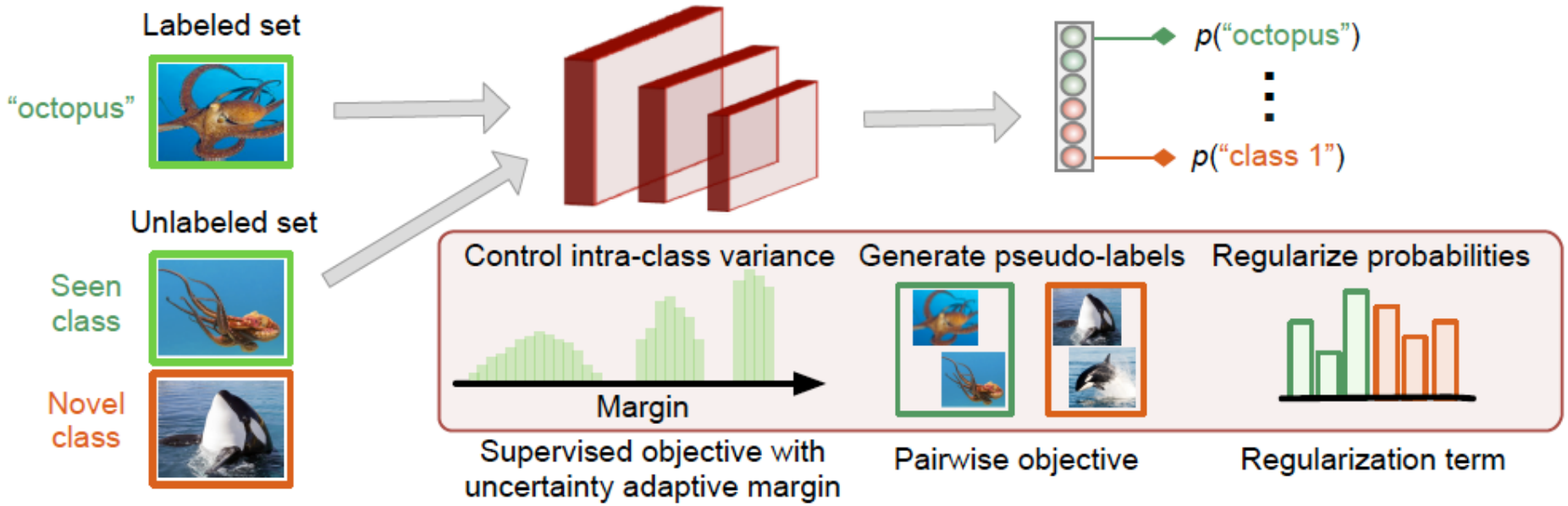


Figure 2: Overview of ORCA framework. ORCA utilizes additional classification heads for novel classes. Objective function in ORCA consists of *(i)* supervised objective with uncertainty adaptive margin, *(ii)* pairwise objective that generates pseudo-labels, and *(iii)* regularization term.

ORCA

Given labeled instances $\mathcal{X}_l = \{x_i \in \mathbb{R}^N\}_{i=1}^n$ and unlabeled instances $\mathcal{X}_u = \{x_i \in \mathbb{R}^N\}_{i=1}^m$, ORCA first applies the embedding function $f_\theta : \mathbb{R}^N \rightarrow \mathbb{R}^D$ to obtain the feature representations $\mathcal{Z}_l = \{z_i \in \mathbb{R}^D\}_{i=1}^n$ and $\mathcal{Z}_u = \{z_i \in \mathbb{R}^D\}_{i=1}^m$ for labeled and unlabeled data, respectively. Here, $z_i = f_\theta(x_i)$ for every instance $x_i \in \mathcal{X}_l \cup \mathcal{X}_u$. On top of the backbone network, we add a classification head consisting of a single linear layer parameterized by a weight matrix $W : \mathbb{R}^D \rightarrow \mathbb{R}^{|\mathcal{C}_l \cup \mathcal{C}_u|}$, and followed by a softmax layer. Note that the number of classification heads is set to the number of previously seen classes and the expected number of novel classes. So, first $|\mathcal{C}_l|$ heads classify instances to one of the previously seen classes, while the remaining heads assign instances to novel classes. The final class/cluster prediction is calculated as $c_i = \text{argmax}(W^T \cdot z_i) \in \mathbb{R}$. If $c_i \notin \mathcal{C}_l$, then x_i belongs to novel classes. The number of novel classes $|\mathcal{C}_u|$ can be known and given as an input to the algorithm which is a typical assumption of clustering and novel class discovery methods. However, if the number of novel classes is not known ahead of time, we can initialize ORCA with a large number of prediction heads/novel classes. The ORCA objective function then infers the number of classes by not assigning any instances to unneeded prediction heads so these heads never activate.

ORCA

However, using standard cross-entropy loss on labeled data creates an imbalance problem between the seen and novel classes, i.e., the gradient is updated for seen classes C_s, but not for novel classes C_n

This can result in learning a classifier with larger magnitudes (Kang et al., 2019) for seen classes, leading the whole model to be biased towards the seen classes.

To overcome the issue an uncertainty adaptive margin mechanism and propose to normalize the logits as we describe next

ORCA

The objective function in ORCA combines three components (Figure 2) *(i)* supervised objective with uncertainty adaptive margin, *(ii)* pairwise objective, and *(iii)* regularization term:

$$\mathcal{L} = \mathcal{L}_{\mathrm{S}} + \eta_1 \mathcal{L}_{\mathrm{P}} + \eta_2 \mathcal{R}, \tag{1}$$

Formula

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Specifically, the supervised objective with uncertainty adaptive margin mechanism is defined as:

$$\mathcal{L}_{\mathrm{S}} = \frac{1}{n} \sum_{z_i \in \mathcal{Z}_l} -\log \frac{e^{s(W_{y_i}^T \cdot z_i + \lambda \bar{u})}}{e^{s(W_{y_i}^T \cdot z_i + \lambda \bar{u})} + \sum_{j \neq i} e^{s W_{y_j}^T \cdot z_i}}, \tag{3}$$

ORCA

The objective function in ORCA combines three components (Figure 2) *(i)* supervised objective with uncertainty adaptive margin, *(ii)* pairwise objective, and *(iii)* regularization term:

$$\mathcal{L} = \mathcal{L}_{\mathrm{S}} + \eta_1 \mathcal{L}_{\mathrm{P}} + \eta_2 \mathcal{R}, \tag{1}$$

Therefore, we only generate pseudo-labels from the most confident positive pairs for each instance within the mini-batch. For feature representations $\mathcal{Z}_l \cup \mathcal{Z}_u$ in a mini-batch, we denote its closest set as $\mathcal{Z}'_l \cup \mathcal{Z}'_u$. Note that $\mathcal{Z}'_l$ is always correct since it is generated using the ground-truth labels. The pairwise objective in ORCA is defined as a modified form of the binary cross-entropy loss (BCE):

$$\mathcal{L}_{\mathrm{P}} = \frac{1}{m+n} \sum_{\substack{z_i, z'_i \in \\ (\mathcal{Z}_l \cup \mathcal{Z}_u, \mathcal{Z}'_l \cup \mathcal{Z}'_u)}} -\log \langle \sigma(W^T \cdot z_i), \sigma(W^T \cdot z'_i) \rangle. \tag{5}$$

Here, σ denotes the softmax function which assigns instances to one of the seen or novel classes.

ORCA

The objective function in ORCA combines three components (Figure 2) *(i)* supervised objective with uncertainty adaptive margin, *(ii)* pairwise objective, and *(iii)* regularization term:

$$\mathcal{L} = \mathcal{L}_{\mathrm{S}} + \eta_1 \mathcal{L}_{\mathrm{P}} + \eta_2 \mathcal{R}, \tag{1}$$

3.5 REGULARIZATION TERM

Finally, the regularization term avoids a trivial solution of assigning all instances to the same class. In early stages of the training, the network could degenerate to a trivial solution in which all instances are assigned to a single class, *i.e.*, $|\mathcal{C}_u| = 1$. We discourage this solution by introducing a Kullback-Leibler (KL) divergence term that regularizes $\Pr(y|x \in \mathcal{D}_l \cup \mathcal{D}_u)$ to be close to a prior probability distribution $\mathcal{P}$ of labels y:

$$\mathcal{R} = KL\Big(\frac{1}{m+n} \sum_{z_i \in \mathcal{Z}_l \cup \mathcal{Z}_u} \sigma(W^T \cdot z_i) \parallel \mathcal{P}(y)\Big), \tag{6}$$

Datasets

** CIFAR-10,*
** CIFAR-100 (Krizhevsky, 2009)*
** ImageNet (Russakovsky et al., 2015),*
** A highly unbalanced single-cell Mouse Ageing Cell Atlas dataset from biology domain (Consortium et al., 2020).*

Results

Table 2: Mean accuracy computed over three runs. Asterisk (∗) denotes that the original method can not recognize seen classes (and we had to extend it). Dagger (†) denotes the original method can not detect novel classes (and we had to extend it). SimCLR and FixMatch are not applicable (NA) to the single-cell dataset. Improvement is computed as a relative improvement over the best baseline.

	CIFAR-10			CIFAR-100			ImageNet-100			Single-cell		
Method	**Seen**	**Novel**	**All**	**Seen**	**Novel**	**All**	**Seen**	**Novel**	**All**	**Seen**	**Novel**	**All**
†FixMatch	71.5	50.4†	49.5	39.6	23.5†	20.3	65.8	36.7†	34.9	NA	NA	NA
†DS^{3}L	77.6	45.3†	40.2	55.1	23.7†	24.0	71.2	32.5†	30.8	76.2	29.7†	26.4
†CGDL	72.3	44.6 †	39.7	49.3	22.5†	23.5	67.3	33.8 †	31.9	74.1	30.4 †	25.6
*DTC	53.9*	39.5	38.3	31.3*	22.9	18.3	25.6*	20.8	21.3	29.6*	25.3	27.8
*RankStats	86.6*	81.0	82.9	36.4*	28.4	23.1	47.3*	28.7	40.3	42.3*	31.9	38.6
*SimCLR	58.3*	63.4	51.7	28.6*	21.1	22.3	39.5*	35.7	36.9	NA	NA	NA
ORCA-ZM	87.6	86.6	86.9	55.2	32.0	34.8	80.4	43.7	55.1	89.5	35.1	47.6
ORCA-FNM	88.0	88.2	88.1	58.2	40.0	44.3	73.0	66.2	68.9	89.7	48.6	58.7
ORCA	**88.2**	**90.4**	**89.7**	**66.9**	**43.0**	**48.1**	**89.1**	**72.1**	**77.8**	**89.9**	**65.2**	**72.9**

Datasets

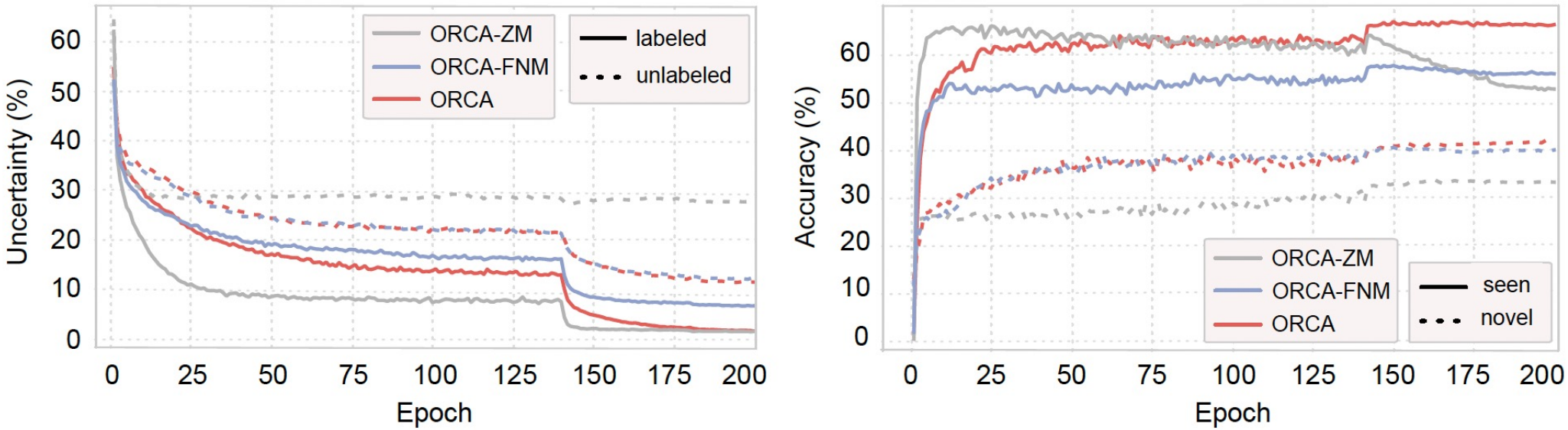


Figure 3: Effect of the uncertainty adaptive margin on the estimated uncertainty (left) and accuracy (right) during training on the CIFAR-100 dataset. At epoch 140, we decay learning rate.

https://abdullah-mamun.com
a.mamun@asu.edu