Rethinking the value of labels for improving class imbalanced learning

NeurIPS 2020

Catalogue

- Background
- Related work
- Introduction
- Method
- Conclusion
- My opinions & questions

Background

Class imbalance

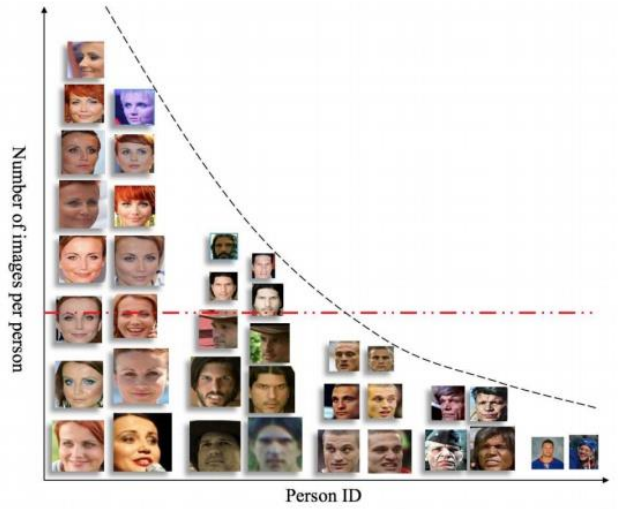
- Privacy issue;
- Low probability of some specific categories…

↓ Thus

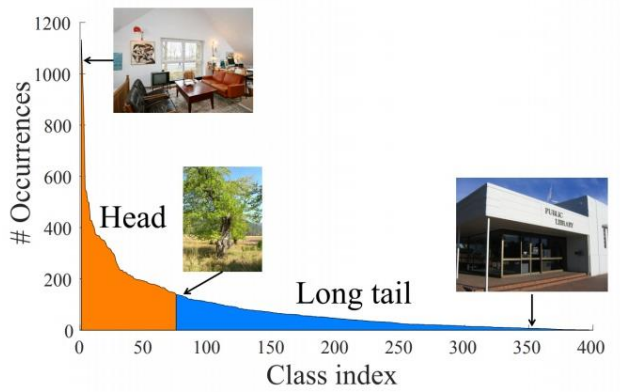
- Difficult to obtain the data of some specific categories
- Large-scale datasets often exhibit long-tailed label distributions

↓ Such as

- Privacy issue in medical diagnosis;
- Low probability of accidents in autonomous driving…



Faces [Zhang et al. 2017]



Places [Wang et al. 2017]

Related work

Imbalanced learning

- Re-sampling(under-sampling, over-sampling)
- Re-weighting(Focal loss, CB loss…)
- Synthetic samples
- Transfer learning
- Metric learning
- Meta learning
- Domain adaption
- Decoupling representation & classifier(SOTA now)

semi-supervised learning (SSL)

self-supervised pre-training (SSP)

Introduction-Value of labels in class-imbalanced learning

- On the positive viewpoint, imbalanced labels are indeed valuable. This is obvious in the semi-supervised learning(SSL).

- On the negative viewpoint, imbalanced labels are not advantageous all the time for label bias. Thus, it’s advisable to learn better initialization that is more label-agnostic from the imbalanced dataset by self-supervised pre-training (SSP) firstly.

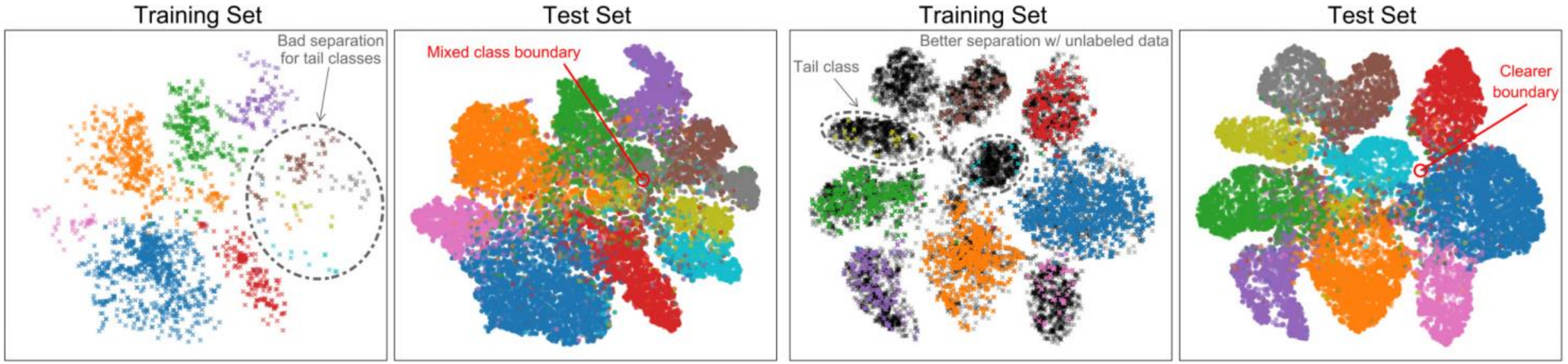
Introduction-Value of labels in class-imbalanced learning

Table 1: Top-1 test errors (%) of ResNet-32 on long-tailed CIFAR-10 and SVHN. We compare SSL using 5x unlabeled data ($\mathcal{D}_U$ @5x) with corresponding supervised baselines. Imbalanced learning can be drastically improved with unlabeled data, which is consistent across different ρ_U and learning strategies.

(a) CIFAR-10-LT

Imbalance Ratio (ρ)	100				50				10			
$\mathcal{D}_U$ Imbalance Ratio (ρ_U)	1	$\rho/2$	ρ	2ρ	1	$\rho/2$	ρ	2ρ	1	$\rho/2$	ρ	2ρ
CE	29.64				25.19				13.61			
CE + $\mathcal{D}_U$ @5x	**17.48**	18.42	18.74	20.06	**16.79**	16.88	18.36	19.94	**10.22**	10.48	10.86	11.04
LDAM-DRW [7]	22.97				19.06				11.84			
LDAM-DRW + $\mathcal{D}_U$ @5x	**14.96**	15.18	15.33	15.55	**14.33**	14.70	14.93	15.24	8.72	**8.24**	8.68	8.97

Introduction-Value of labels in class-imbalanced learning



(a) Standard CE training (b) With unlabeled data $\mathcal{D}_U$ @5x (colored in black)

Figure 1: t-SNE visualization of training & test set on SVHN-LT. Using unlabeled data helps to shape clearer class boundaries and results in better class separation, especially for the tail classes.

Method-SSL-Theoretical Motivation

The settings of abstract model:

- Binary classification problem;
- The data distribution P_{XY} being a mixture of two Gaussians.
- The label Y is either positive (+1) or negative (-1) with equal $\mathrm{P} = 0.5$
- Distribution of X: $\begin{cases} \mathrm{X} \sim \mathrm{N}(\mu_1, \sigma^2), \ \mathrm{Y} = 1 \\ \mathrm{X} \sim \mathrm{N}(\mu_2, \sigma^2), \ \mathrm{Y} = -1 \end{cases}$

- The optimal Bayes's classifier: $\mathrm{f}(\mathrm{x}) = \mathrm{sign}\left(\mathrm{x} - \frac{\mu_1 + \mu_2}{2}\right), \ \mathrm{x} \in \mathrm{X}, \ \mathrm{f}(\mathrm{x}) \in \mathrm{Y}$
- Extra unlabled data(hypothesis: iid): $\tilde{\mathrm{X}}$
- The number of positive(negative) sample in $\tilde{\mathrm{X}}$: $\tilde{\mathrm{n}}_+ \left(\tilde{\mathrm{n}}_-\right)$
- Δ can be seen as the representation of imbalanceness

Method-SSL-Theoretical Motivation

$\frac{\mu_1+\mu_2}{2}$ can be seen as the representation of imbalanceness

$$\hat{\theta} = \frac{1}{2}\left(\sum_{i=1}^{\tilde{n}_+} \tilde{X}_i^+ / \tilde{n}_+ + \sum_{i=1}^{\tilde{n}_-} \tilde{X}_i^- / \tilde{n}_-\right)$$

Theorem 1. *Consider the above setup. For any* $\delta > 0$, *with probability at least* $1 - 2e^{-\frac{2\delta^2}{9\sigma^2}\cdot\frac{\tilde{n}_+\tilde{n}_-}{\tilde{n}_-+\tilde{n}_+}} - 2e^{-\frac{8\tilde{n}_+\delta^2}{9(\mu_1-\mu_2)^2}} - 2e^{-\frac{8\tilde{n}_-\delta^2}{9(\mu_1-\mu_2)^2}}$ *our estimates* $\hat{\theta}$ *satisfies*

$$\left|\hat{\theta} - (\mu_1+\mu_2)/2 - \Delta(\mu_1-\mu_2)/2\right| \leq \delta.$$

The results based on theorem 1:

- Training data imbalance affects the accuracy of our estimation.
- Unlabeled data imbalance affects the probability of obtaining such a good estimation.

Method-SSL-Framework & notes

Classic self-training framework: $\mathcal{L}(\mathcal{D}_L, \theta) + \omega\mathcal{L}(\mathcal{D}_U, \theta)$

Other SSL framework: besides self-training, more advanced SSL techniques can be easily incorporated into our framework by modifying only the loss function.

Notes:

- If D_U is more balanced, the gains resulting from D_U are larger.
- If D_L is more balanced, the gains resulting from different D_U are similar.
- Larger D_U or D_L often brings higher gains, with gains gradually diminish as data amount grows.
- SSL can lead to consistent gains across all classes, where trends are more evident for tail classes

Method-SSL-Notes & part of experiment

Notes:

- The relevance between the unlabeled data and the original data has to be as high as 60% to be effective.
- In this case, to be helpful, ρ_U(the imbalanceness of the unlabled data) cannot be larger than 50 (which is the imbalance ratio of the original data)

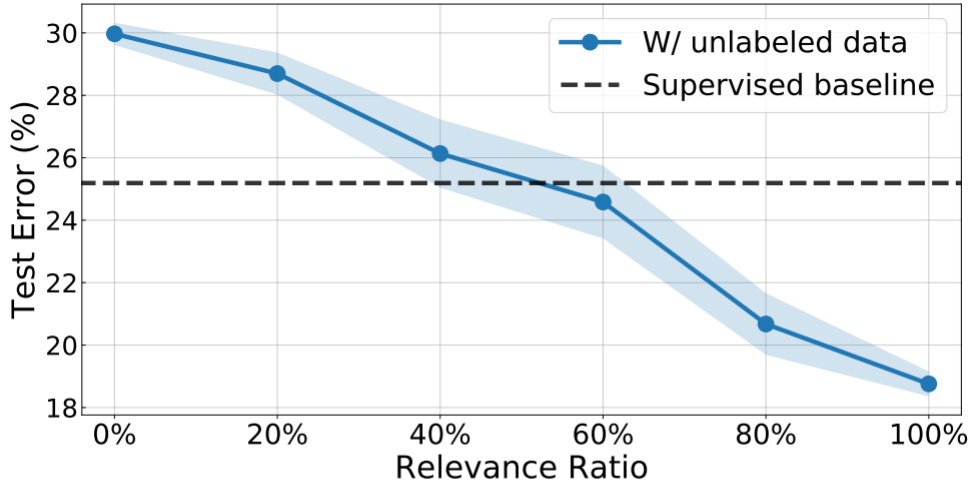


Figure 2: Test errors of different unlabeled data relevance ratios on CIFAR-10-LT with $\rho = 50$. We fix $\rho_U = 50$ for the relevant unlabeled data.

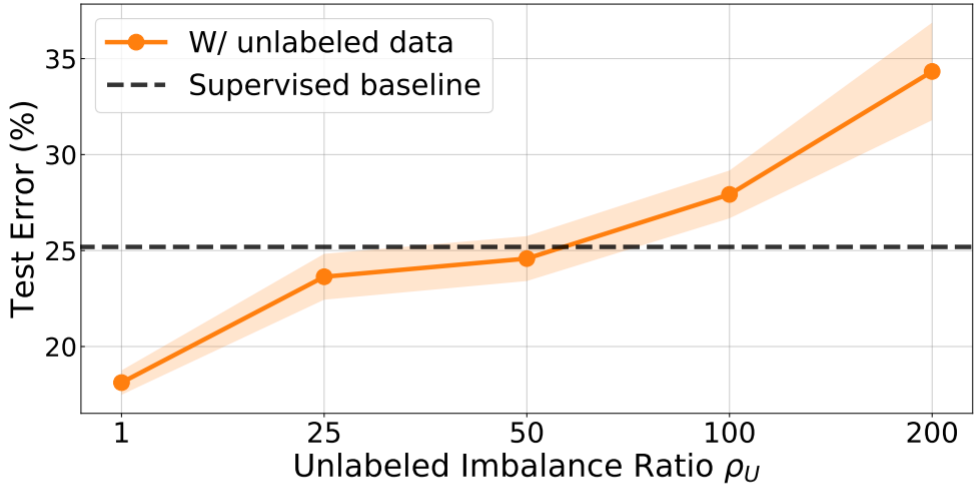


Figure 3: Test errors of different ρ_U of relevant unlabeled data on CIFAR-10-LT with $\rho = 50$. We fix the unlabeled data relevance ratio as 60%.

Method-(Self-Supervision)-Theoretical Motivation

The settings of abstract model:

- Binary classification problem;
- The data distribution P_{XY} being a mixture of d-dimensional Gaussians.
- The label Y is either positive (+1) or negative (-1) with $p_- > p_+$
- Distribution of X: $\begin{cases} \mathrm{X} \sim \mathrm{N}(0, {\sigma_1}^2 \mathrm{I_d}), \ \mathrm{Y} = 1 \\ \mathrm{X} \sim \mathrm{N}(\mu_2, \beta {\sigma_2}^2 \mathrm{I_d}), \ \mathrm{Y} = -1 \end{cases}$

- The linear classifier: $\mathrm{f}(\mathrm{x}) = \mathrm{sign}(< \theta, \mathrm{x} > + \mathrm{b}), \ \mathrm{x} \in \mathrm{X}, \ \mathrm{f}(\mathrm{x}) \in \mathrm{Y}$
- The learned representation of SS: $Z = \psi(\mathrm{X}) = k_1 ||X||_2^2 + k_2, \ where \ k_1, k_2 > 0$

Method-(Self-Supervision)-Theoretical Motivation

$$err(\mathrm{f}) = \sum_{P(X,Y)} \mathrm{p}(\mathrm{f}(\mathrm{x}) \neq \mathrm{y}) \quad f_{ss}(X) = \mathrm{sign}(-Z + b),\ b = \frac{1}{2}\left(\frac{\sum_{i=1}^{N} \mathbf{1}_{\{Y_i=+1\}} Z_i}{N_+} + \frac{\sum_{i=1}^{N} \mathbf{1}_{\{Y_i=-1\}} Z_i}{N_-}\right)$$

Theorem 2. *Let Φ be the CDF of $\mathcal{N}(0,1)$. For any linear classifier of the form $f(X) = sign(\langle\theta, X\rangle + b)$ where $b > 0$, the error probability satisfies:* $\mathrm{err}_f = p_+\Phi\left(-\frac{b}{||\theta||_2\sigma_1}\right) + p_-\Phi\left(\frac{b}{||\theta||_2\sqrt{\beta}\sigma_1}\right) \geq \frac{1}{4}$.

Theorem 3. *Consider the linear classifier with self-supervised learning, f_{ss}. For any $\delta \in \left(0, \frac{\beta-1}{\beta+1}\right)$, we have that with probability at least $1 - 2e^{-N_- d\delta^2/8} - 2e^{-N_+ d\delta^2/8}$, the classifier satisfies*

$$\mathrm{err}_{f_{ss}} \leq \begin{cases} p_+ e^{-d\cdot\frac{(\beta-1-(1+\beta)\delta)^2}{32}} + p_- e^{-d\cdot\frac{(\beta-1-(1+\beta)\delta)^2}{32\beta^2}}, & \text{if } \delta \in \left[\frac{\beta-3}{\beta+1}, \frac{\beta-1}{\beta+1}\right); \\ p_+ e^{-d\cdot\frac{(\beta-1-(1+\beta)\delta)}{16}} + p_- e^{-d\cdot\frac{(\beta-1-(1+\beta)\delta)^2}{32\beta^2}}, & \text{if } \delta \in \left(0, \frac{\beta-3}{\beta+1}\right). \end{cases}$$

The results based on theorems above:

- With high probability, we obtain a satisfying classifier f_{ss}, whose error probability decays exponentially on the dimension d.
- Training data imbalance(i.e. β) affects our probability of obtaining such a satisfying classifier.

Method-(Self-Supervision)-Framework & notes

Self-supervised pre-training (SSP): After the first stage of learning with self-supervision, we can then perform any standard training approach to learn the final model initialized by the pre-trained network.

Other SSP methods: All SSP methods can lead to notable gains compared to the baseline, while interestingly the gain varies across methods.

Notes:

- If the original data is more imbalanced, the gains resulting from the self -supervision are larger.
- Regardless of the settings and the base training techniques, adding our self-supervision framework in the first stage of learning can uniformly boost the final performance.
- SSP can lead to consistent gains across all classes, where trends are more evident for tail classes

Conclusion

Two novel perspectives:

- (1)using unlabeled data without depending on additional human labeling;
- (2)explore intrinsic properties from data itself with self-supervision.

Challenges in real world

- Some ethical issues like privacy and fairness may impose additional constraints in learning process and results.

My opinion & questions about the paper

Opinion:

The two perspectives is really novel and useful, but I think they will still have many challenges when used in imbalanced learning. For example, in SSL, the information of unlabled data is unknown, so for many imbalanced learning application, there are a high probability of obtaining a bad extra data.

All in all, I think self-supervision may be better than semi-supervision when it comes to convenience and constraints.

Question:

- What the result will be if we combined the SSL and SSP?
- Where is the quantitative description of the prerequisites about the original data when using the self-supervision pre-training?

Thanks for listening.