南京航空航天大學
Nanjing University of Aeronautics and Astronautics

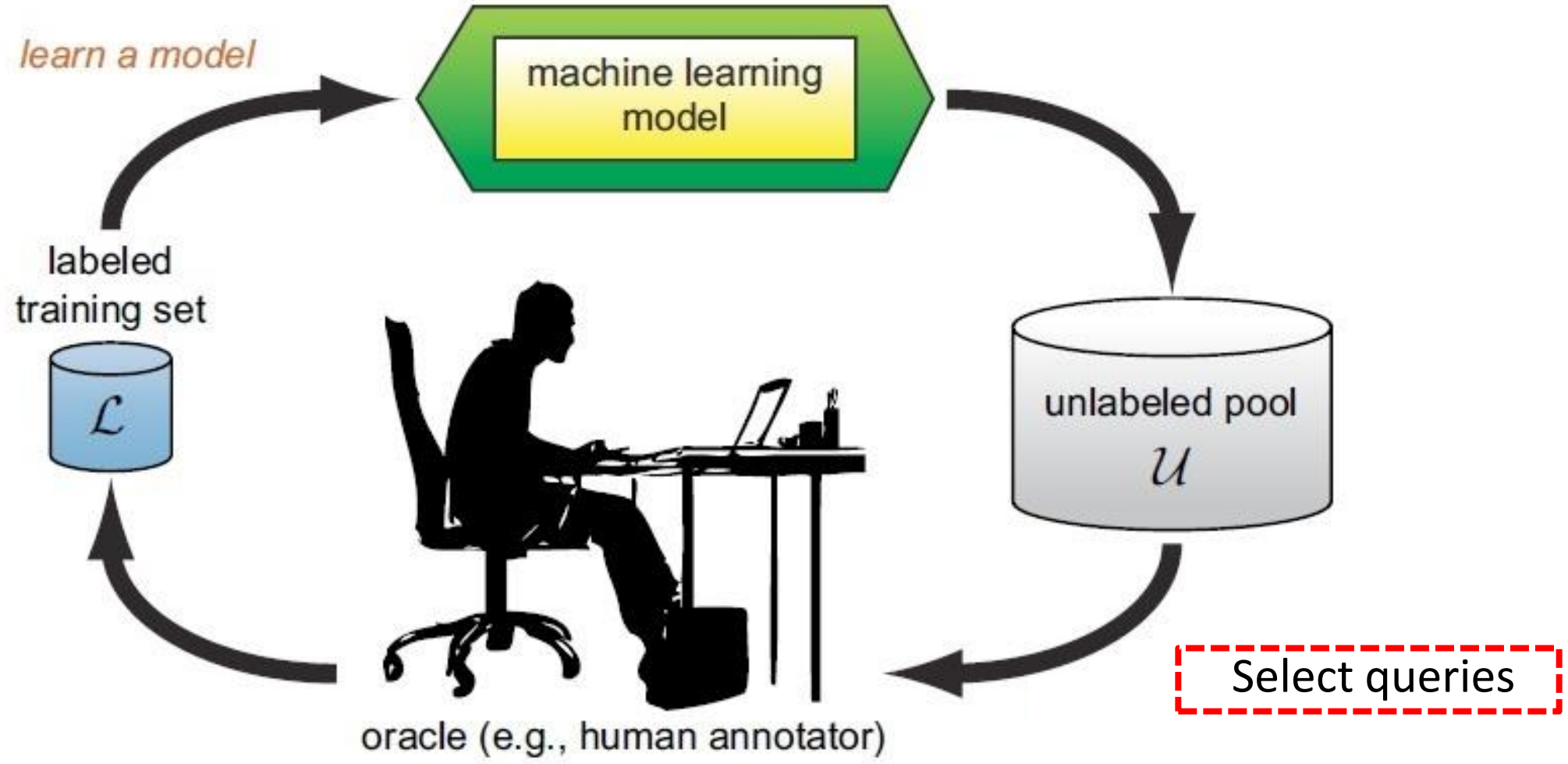
VaB-AL： Incorporating Class Imbalance and Difficulty with Variational Bayes for Active Learning

Jongwon Choi[1*], Kwang Moo Yi[2*], Jihoon Kim[3], Jinho Choo[3],
Byoungjip Kim[3], Jinyeop Chang[3], Youngjune Gwon[3], Hyung Jin Chang[4]

[1]Chung-Ang University, South Korea
[2]University of British Columbia, Canada
[3]Samsung SDS, South Korea
[4]University of Birmingham, United Kingdom

CVPR 2021

Active Learning



How to select useful samples ?

Active learning based on the hardest examples

acquire those that have the highest probability of making wrong predictions.

$$p(y \neq \widehat{y}|\mathbf{x})$$

Not a probability distribution that can be modelled directly

N_c : the number of classes

$p(y_n)$: $p(y = n)$

$$p(y \neq \widehat{y}|\mathbf{x}) = 1 - p(y = \widehat{y}|\mathbf{x}) = 1 - \sum_{n=1}^{N_c} p(y_n, \widehat{y}_n|\mathbf{x})$$

Bayesian active learning

$$p(y_n, \widehat{y}_n|\mathbf{x}) = \frac{p(y_n|\widehat{y}_n, \mathbf{x})p(\mathbf{x}|\widehat{y}_n)p(\widehat{y}_n)}{\sum_{n=1}^{N_c} p(\mathbf{x}|\widehat{y}_n)p(\widehat{y}_n)}$$

we assume that the probability of a model making a mistake is highly related to the label.

$$p(y_n, \widehat{y}_n|\mathbf{x}) \approx \frac{p(y_n|\widehat{y}_n)p(\mathbf{x}|\widehat{y}_n)p(\widehat{y}_n)}{\sum_{n=1}^{N_c} p(\mathbf{x}|\widehat{y}_n)p(\widehat{y}_n)}.$$

$$p(y \neq \widehat{y}|\mathbf{x}) \approx 1 - \sum_{n=1}^{N_c} \frac{p(y_n|\widehat{y}_n)p(\mathbf{x}|\widehat{y}_n)p(\widehat{y}_n)}{\sum_{n=1}^{N_c} p(\mathbf{x}|\widehat{y}_n)p(\widehat{y}_n)}.$$

1. the probability of a model making mistake based on label

2. the likelihood of a sample given predicted label

3. the prior on the distribution of predicted labels

Represents how imbalanced the predictions of a model are.

Estimating probabilities with regularized VAE (1)

Likelihood of a sample – $p(\mathbf{x}|\widehat{y}_n)$

In more detail, if we denote the j-th embedding dimension of VAE as z_j and write $j \in C_n$ to denote dimension j is related to class n, we write this absence condition as

$$\widehat{y} = \underset{n}{\operatorname{argmin}} \left[\sum_{j \in C_1} z_j^2, \sum_{j \in C_2} z_j^2, \ldots, \sum_{j \in C_{N_c}} z_j^2 \right]^\top . \quad (6)$$

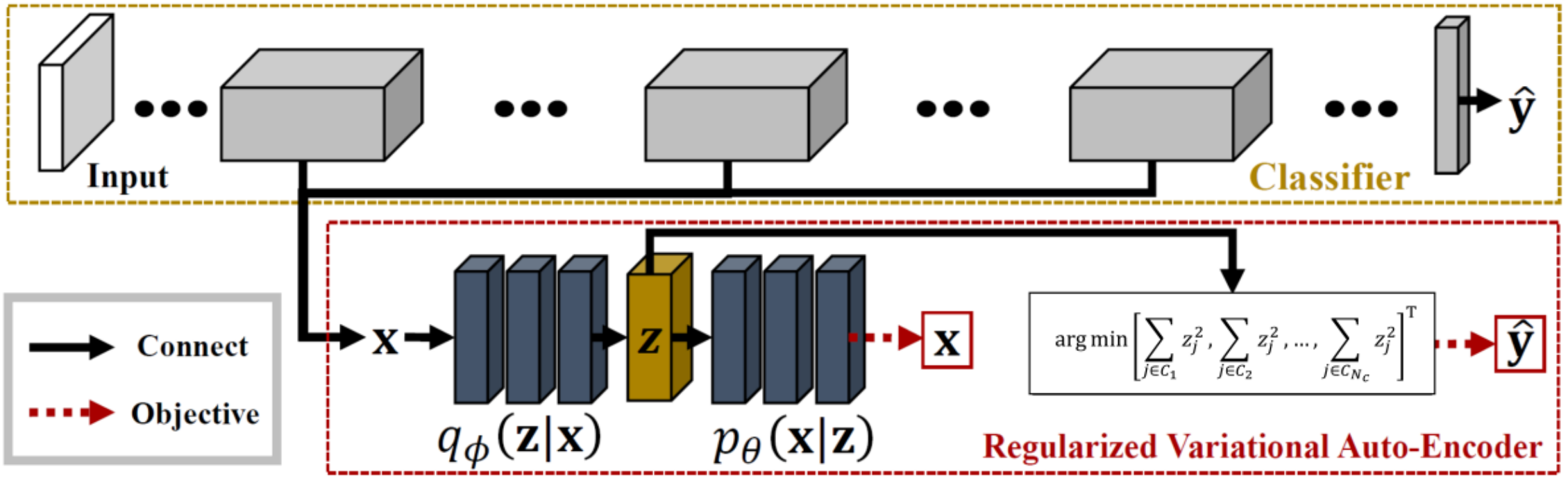
Let $\mathbf{w} = [w_1, w_2, \ldots, w_n]^\top$, where $w_n = \sum_{j \in C_n} z_j^2$

$$\mathcal{L}_{\text{Class}} = \mathcal{H}\left(\text{softmax}\left(-\mathbf{w}\right), \widehat{\mathbf{y}}\right)$$

The empirical lower bound (ELBO)

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + D_{KL}\left(q_\phi\left(\mathbf{z}|\mathbf{x}\right) \parallel p(\mathbf{z})\right) \quad \Rightarrow \quad p(\mathbf{x}|\hat{y}_n)$$

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \lambda\mathcal{L}_{\text{Class}}$$



Estimating probabilities with regularized VAE (2)

Probability of labelling error – $p(y_n|\widehat{y}_n)$.

We use the labels given by the VAE for the data samples in the labelled pool $\mathcal{P}_L$

$$p(y_n|\widehat{y}_n) \approx \frac{\mathbb{E}_{\mathbf{x}\in\mathcal{P}_L,\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\delta\left(y^{(i)},\widehat{y}^{(i)},n\right)\right]}{\mathbb{E}_{\mathbf{x}\in\mathcal{P}_L,\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\delta\left(\widehat{y}^{(i)},n\right)\right]}$$

an indicator function δ : 1 if all inputs are equal and 0 otherwise

Prior – $p(\widehat{y}_n)$

$$p(\widehat{y}_n) \approx \mathbb{E}_{\mathbf{x}\in\mathcal{P}_L,\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\delta\left(\widehat{y}^{(i)},n\right)\right]$$

Algorithm

Algorithm 1: Proposed Method

Input: $\mathcal{P}_U^{(0)}, \mathcal{P}_L^{(0)}, N_r, \mathcal{M}$

$r = 0$ **while** *not at the maximum round* **do**

- Train $\mathcal{M}$ using $\mathcal{P}_L^{(r)}$
- Freeze $\mathcal{M}$
- Train the VAE module using $\mathcal{P}_U^{(0)}$ by Eq. (9)
- Estimate $p(\mathbf{x}|\widehat{y}_n)$ by Eq. (8)
- Estimate $p(y_n|\widehat{y}_n)$ by Eq. (10)
- Estimate $p(\widehat{y}_n)$ by Eq. (11)
- Estimate $p(y \neq \widehat{y}|\mathbf{x})$ by Eq. (2)
- $\mathcal{X} \leftarrow N_r$ samples with the highest uncertainty
- $\mathcal{P}_U^{(r+1)} \leftarrow \mathcal{P}_U^{(r)} - \mathcal{X}$
- $\mathcal{P}_L^{(r+1)} \leftarrow \mathcal{P}_L^{(r)} \bigcup \mathcal{X}$
- $r \leftarrow r + 1$

end

$$p(y_n|\widehat{y}_n) \approx \frac{\mathbb{E}_{\mathbf{x}\in\mathcal{P}_L,\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\delta\left(y^{(i)},\widehat{y}^{(i)},n\right)\right]}{\mathbb{E}_{\mathbf{x}\in\mathcal{P}_L,\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\delta\left(\widehat{y}^{(i)},n\right)\right]}.$$

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + D_{KL}\left(q_\phi\left(\mathbf{z}|\mathbf{x}\right) \parallel p(\mathbf{z})\right)$$

$$p(\widehat{y}_n) \approx \mathbb{E}_{\mathbf{x}\in\mathcal{P}_L,\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\delta\left(\widehat{y}^{(i)},n\right)\right]$$

$$p(y \neq \widehat{y}|\mathbf{x}) = 1 - p(y = \widehat{y}|\mathbf{x}) = 1 - \sum_{n=1}^{N_c} p(y_n, \widehat{y}_n|\mathbf{x}),$$

Experimental Results

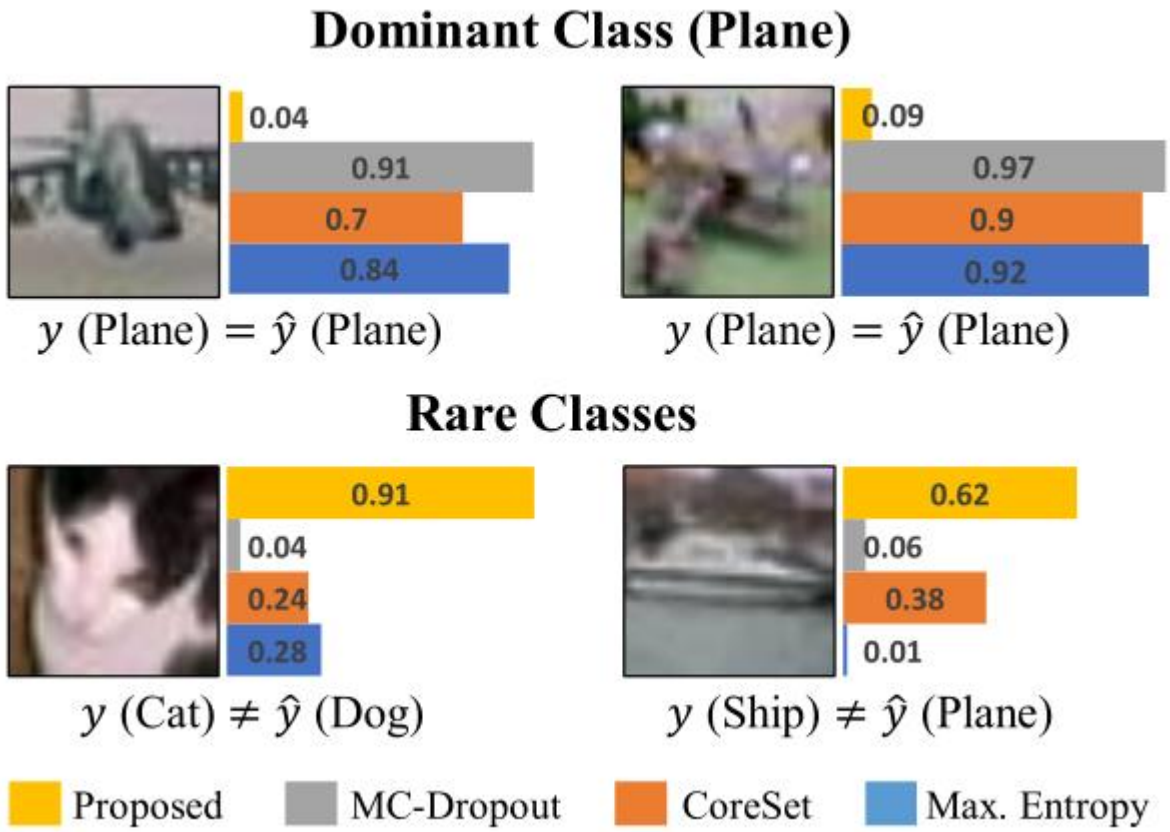


Figure 1. **Class matters**

Experimental Results

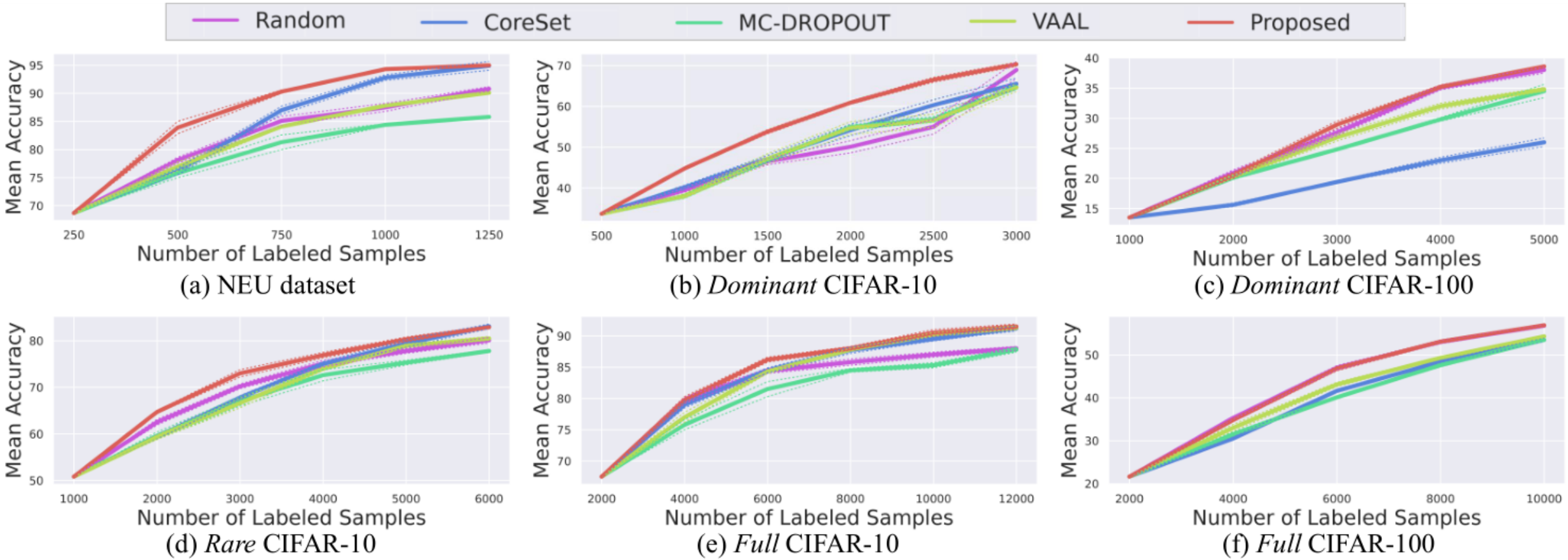


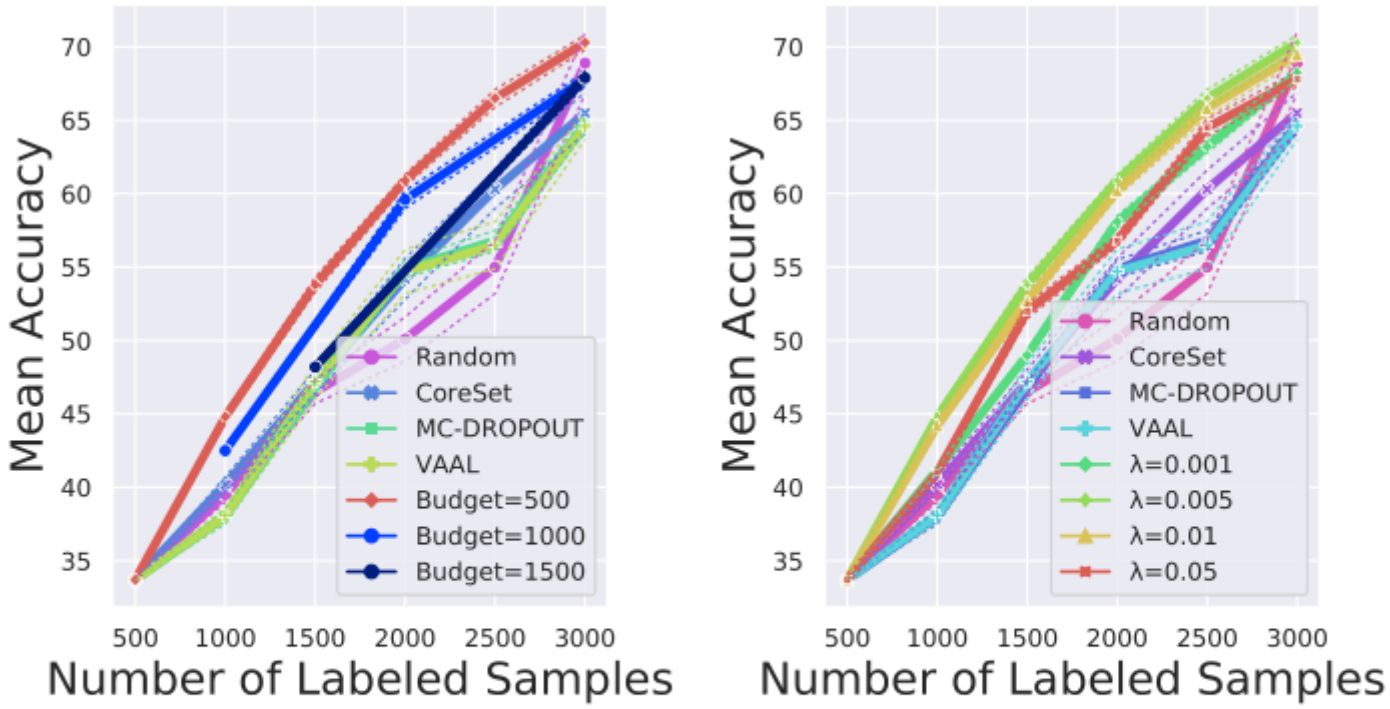
Figure 3. Results for the dominant, rare, and full datasets.

Experimental Results

$p(\widehat{y})$	$p(y\|\widehat{y})$	*CIFAR-10*		*CIFAR-10*$^{+[1]}$	
		avg.	final	avg.	final
-	-	82.56%	88.66%	48.02%	60.13%
✓	-	82.91%	90.68%	**54.23%**	70.25%
-	✓	82.71%	90.51%	51.86%	65.98%
✓	✓	**83.36%**	**91.12%**	54.16%	**70.35%**

Table 1. Validity of the prior $p(\widehat{y})$ and label difficulty $p(y|\widehat{y})$.

Experimental Results



(a) Varying budget sizes. (b) Varying λ.

Figure 5. Results of ablation tests on *dominant* CIFAR-10.

Conclusion

We have proposed a novel active learning method that incorporates class imbalance and label difficulty.

We have shown that this creates a significant difference for a real-world dataset that exhibits data imbalance, as well as in cases when data imbalance is introduced to CIFAR-10 and CIFAR-100 datasets.

THANKS