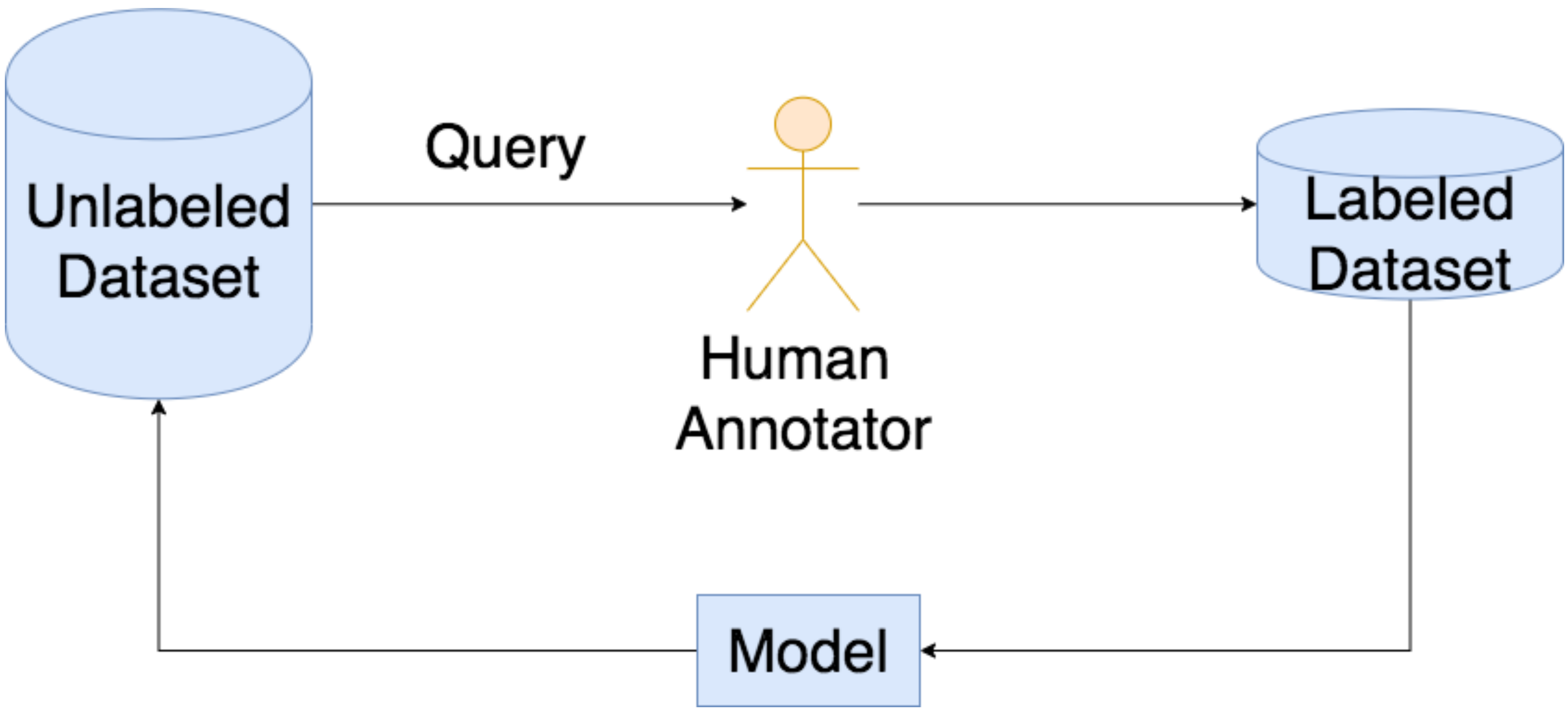
Paper Id-*2743* *ICPR 2020*

Minority Class Oriented Active Learning for Imbalanced Datasets

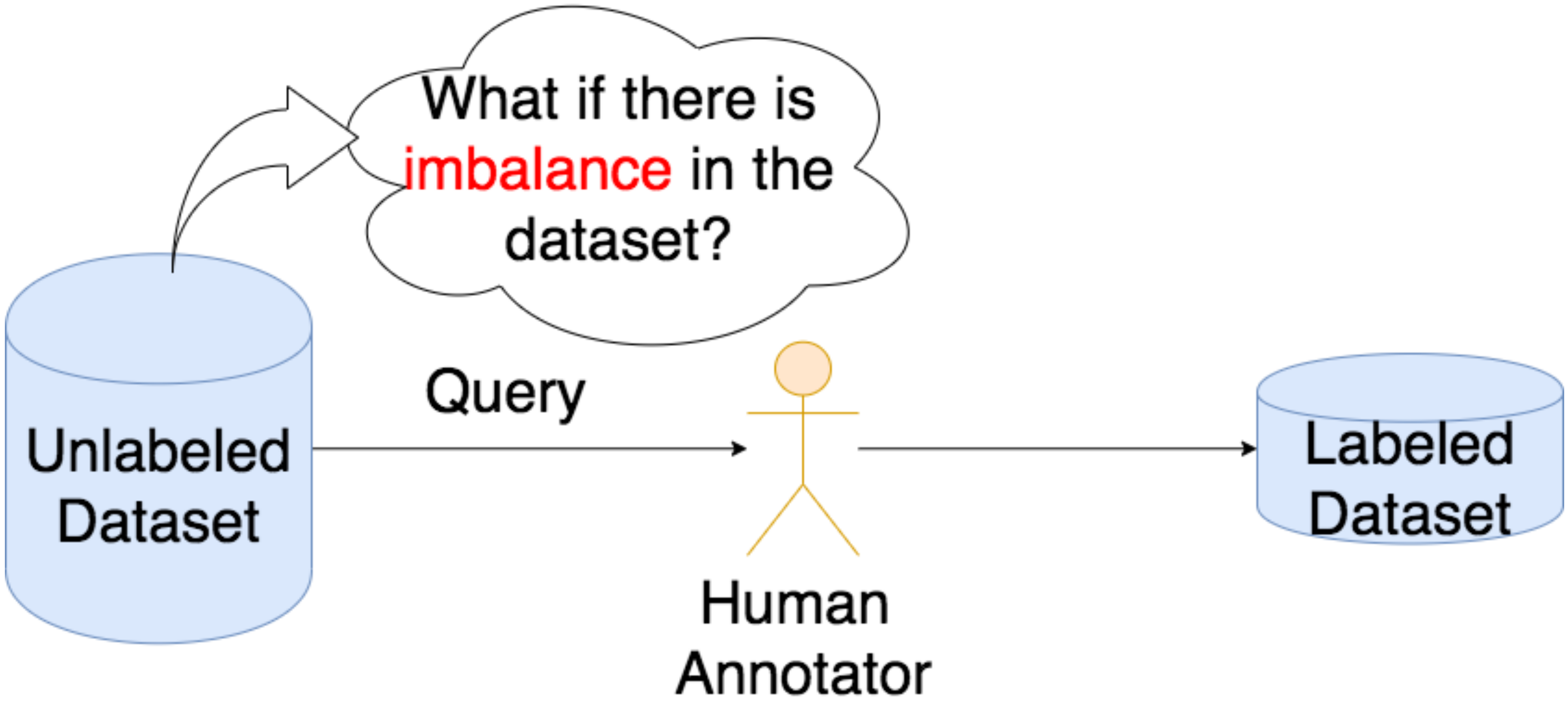
Umang AGGARWAL[1,2], Adrian POPESCU[1], Celine HUDELOT[2]

[1]Université Paris-Saclay, CEA, Département Intelligence Ambiante et Systèmes Interactifs
[2]Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes
umang.aggarwal, adrian.popescu@cea.fr, celine.hudelot@centralesupelec.fr

Iterative Active learning Cycle

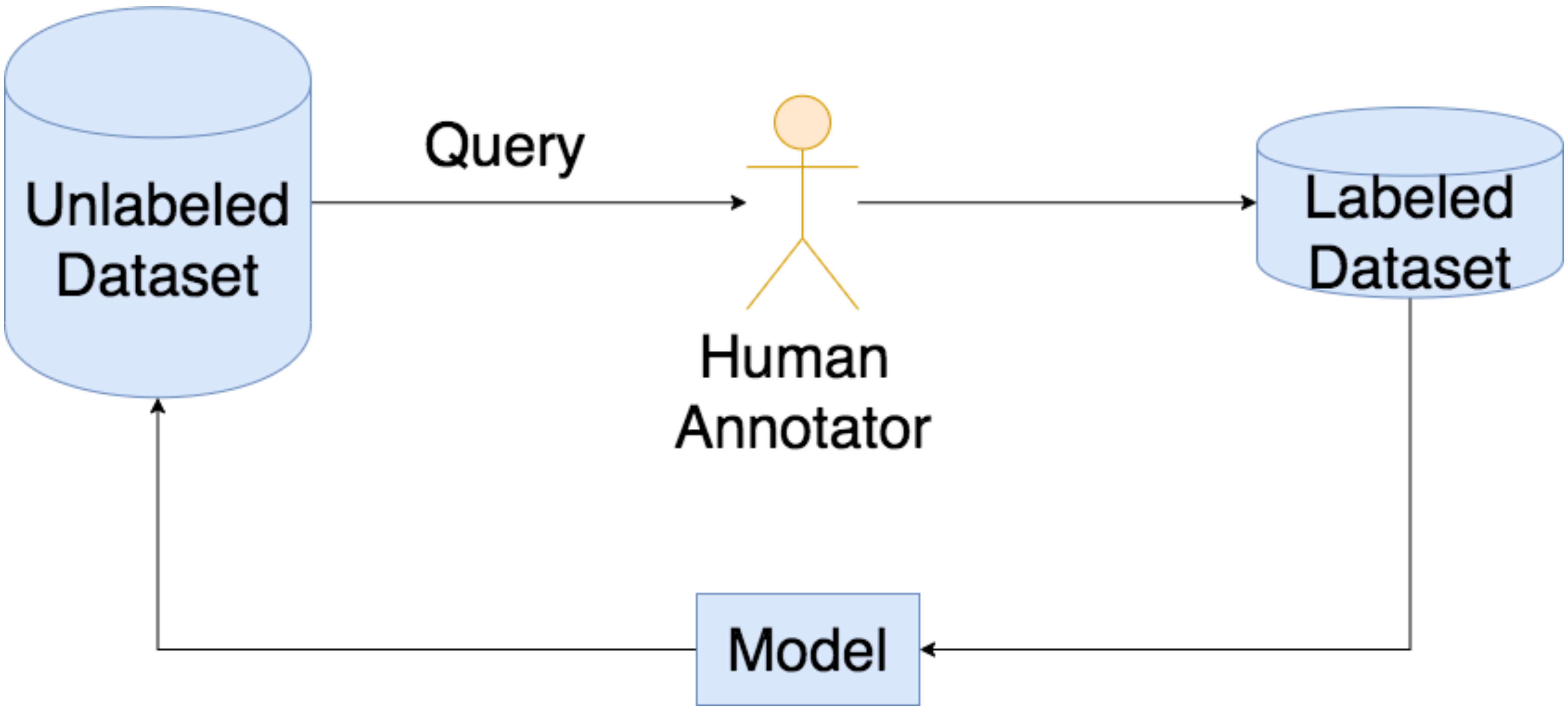


Context


What if there is imbalance in the dataset?
Unlabeled Dataset
Query
Human Annotator
Labeled Dataset

Motivation

Active learning as a way to make informative , diverse and balanced selection over unlabelled dataset.



Iterative Active learning Cycle

Selection criteria

Informativeness

1. Uncertainty based:
   - Least confidence
   - Margin sampling
   - Max entropy

2. Query-by-committee- multiple classifiers

3. Expected model change - loss gradient

Representativeness

1. Clustering based approaches

2. Farthest-first traversal

3. CoreSet

Proposed Method

Minority Class Oriented Sampling

1) Selecting samples predicted as minority class

Samples selected for a class:

$$\mathbb{D}_c^{U(k)} = \{\forall x \in \mathbb{D}_k^U, if\ P(c^1 = c|x)\}$$

Motivation:

if the sample is annotated as minority class :

help to mitigate imbalance

else if annotated as majority class :

help in decision boundary of minority class

2) Number of samples per class depends on imbalance and budget

For a given class (c), at iterative step (k):

Average number of class (μ_k) - Budget / number of classes.

Number of samples in class (c) - s_k^c;

$$m_k^c = \begin{cases} \mu_k - s_k^c, & \text{if } s_k^c < \mu_k \\ 0, & \text{otherwise} \end{cases}$$

3) Allows use of any other AF if imbalance is mitigated or if not enough minority class samples for found

Proposed Method

1) Certainty-oriented Minority Class Sampling

$$CMCS = arginvsort_{\forall x \in \mathbb{D}_c^{U(k)}} marg(x)$$

2) Uncertainty-oriented Minority Class Sampling

$$UMCS = argsort_{\forall x \in \mathbb{D}_c^{U(k)}} marg(x)$$

3)Diversity-oriented Minority Class Sampling

$$DMCS = core(\mathbb{D}_c^{U(k)}, \mathbb{D}_c^{L(k)})$$

Experimental setup

Dataset	Class	Images	Mean(μ)	Std(σ)	ir
FOOD-101	101	22956	227.28	180.31	0.793
CIFAR-100	100	17168	171.68	126.98	0.740
MIT-67	67	14281	213.15	168.16	0.789

TABLE I
DATASET STATISTICS. ir IS THE IMBALANCE RATIO.

Initial Budget- 500

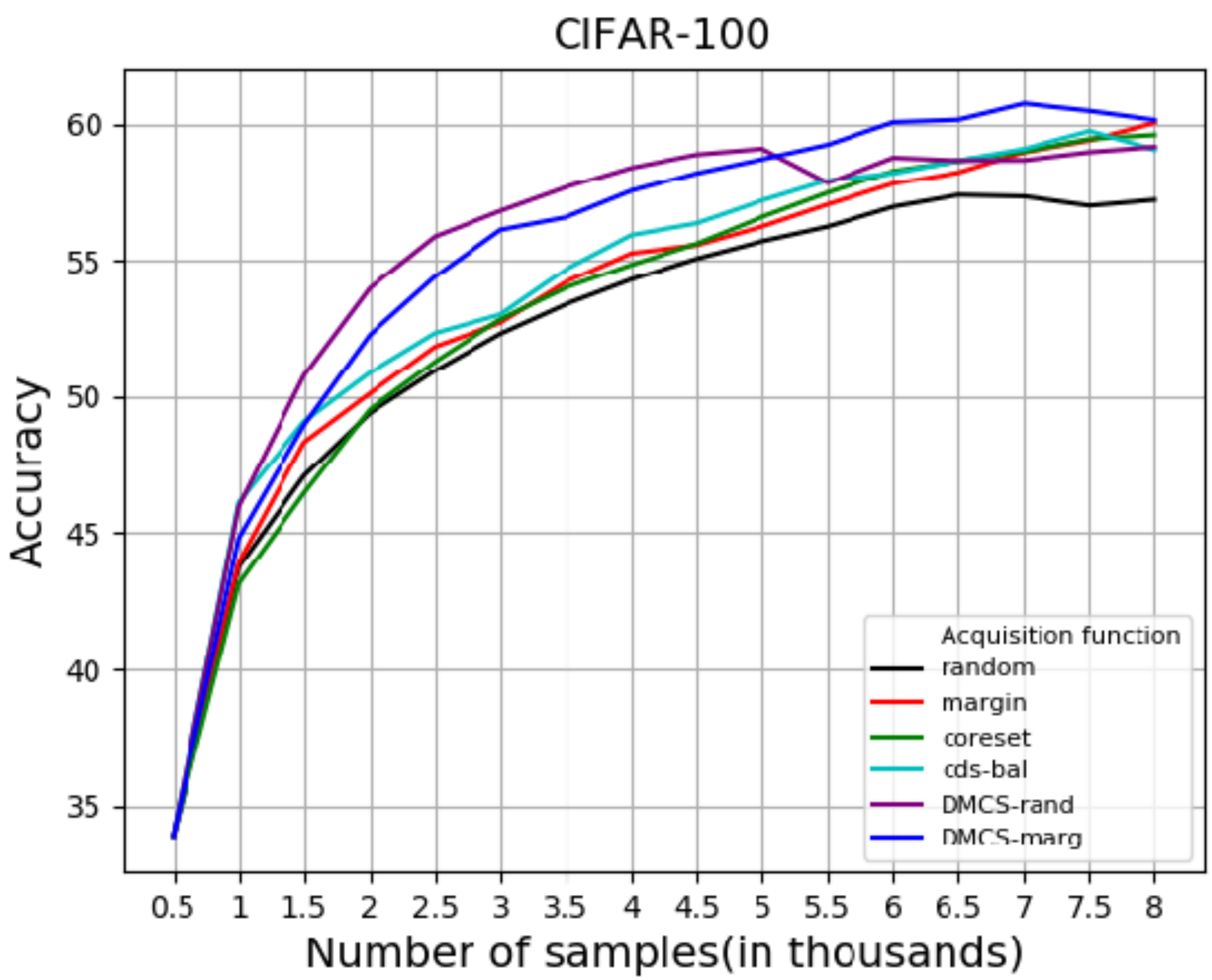
Iteration- 15, Total budget - 8000

Model- ResNet18

Training schemes

1. Fine-tuning ResNet18 with thresholding
2. Cost-Sensitive SVM over pre-trained ResNet18 features

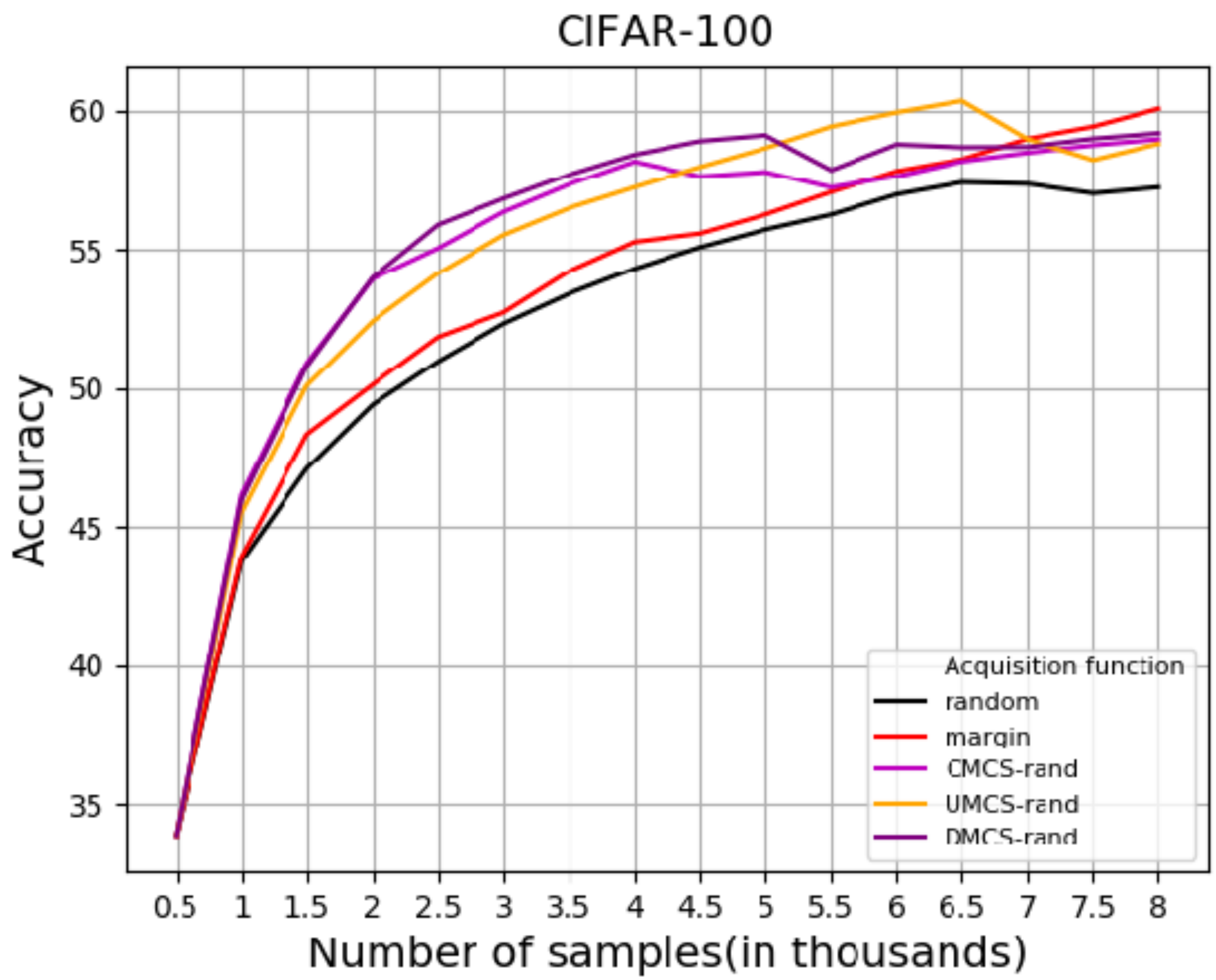
Results



Iterative active learning performance for baselines and for the proposed method DMCS

The AL budget is 8000 and the number of iterations is 15

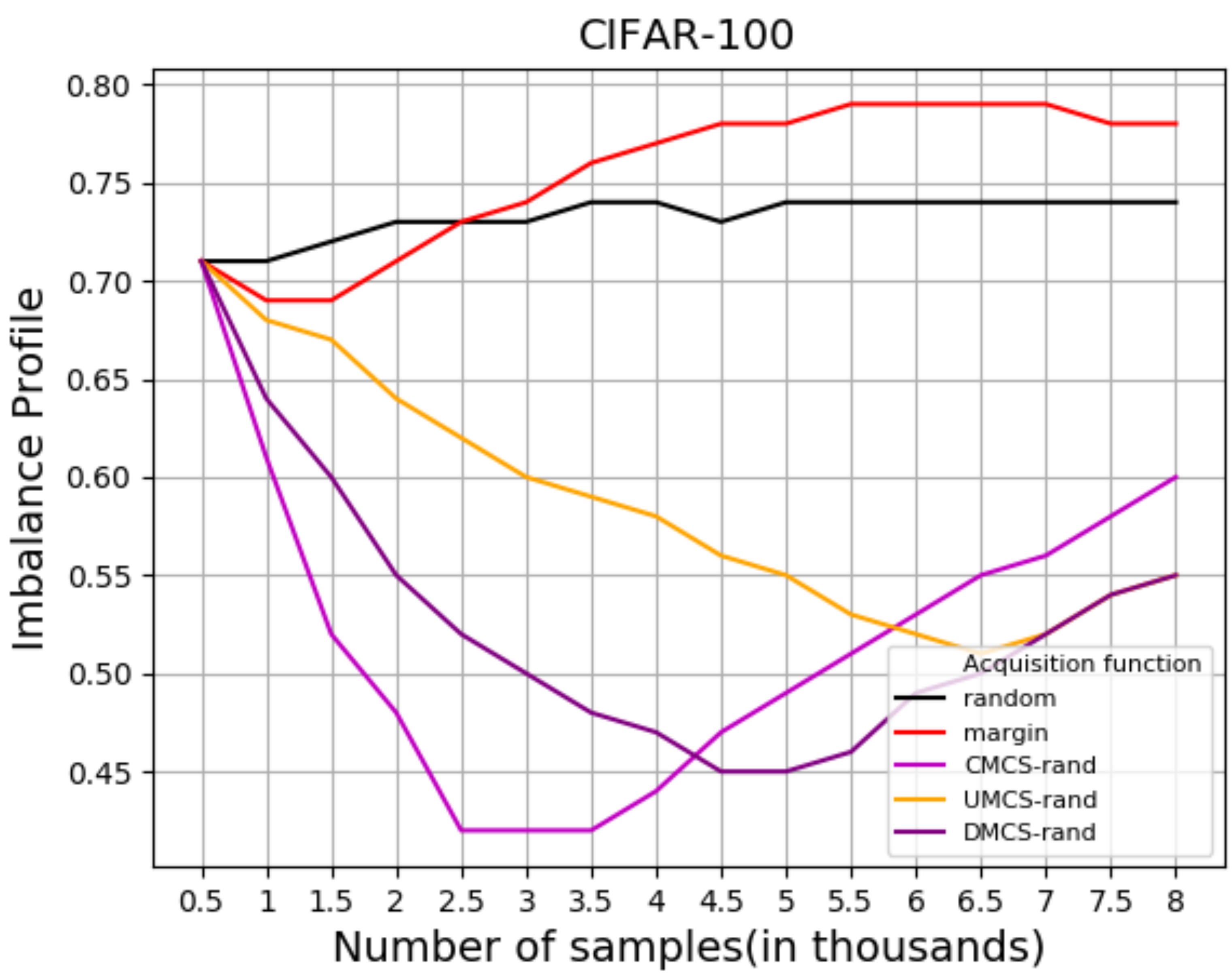
Results



Iterative active learning performance for baselines and three variants of the proposed method.

The AL budget is 8000 and the number of iterations is 15.

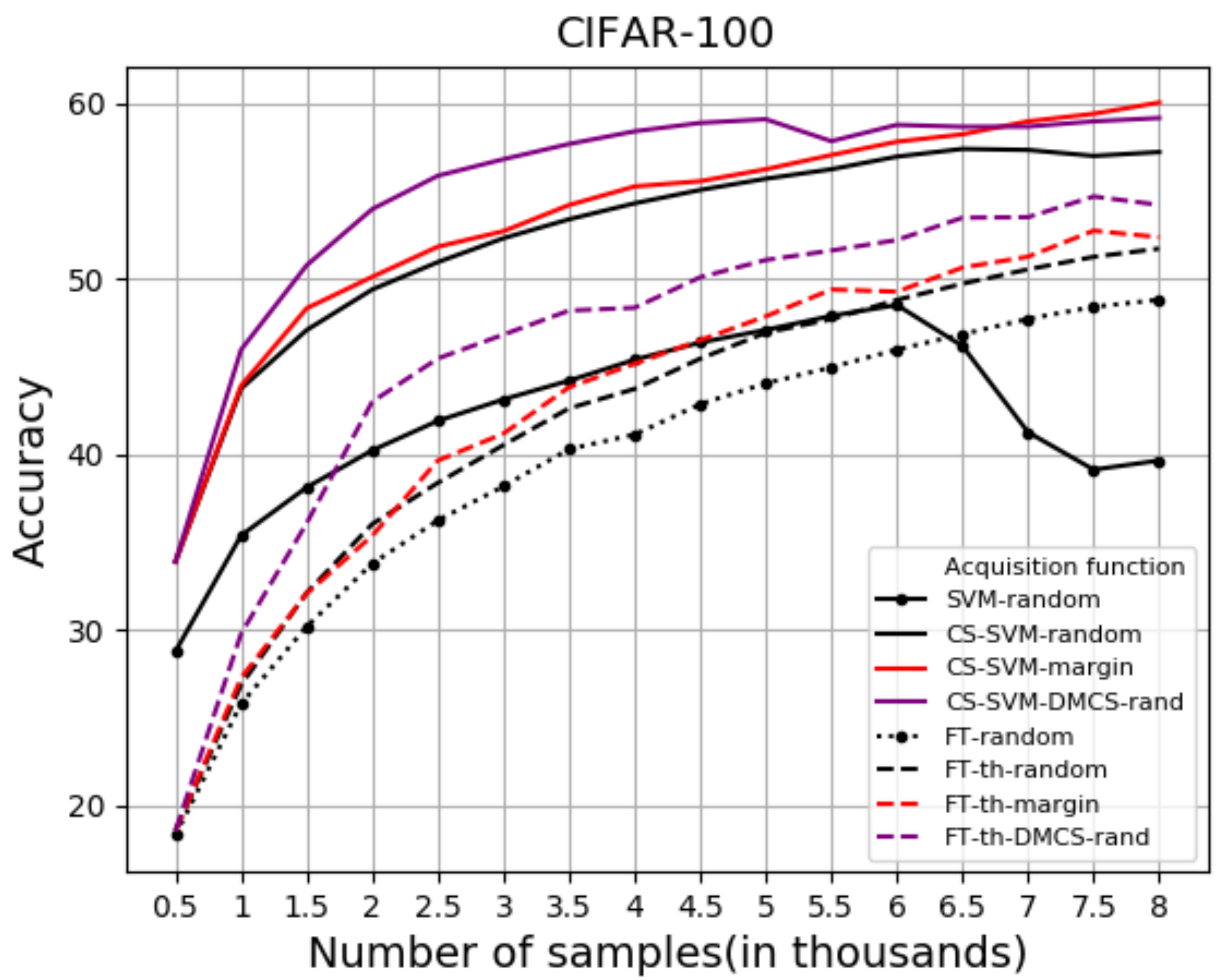
Results



Imbalance Profiles for baselines and three variants of the proposed method.

The AL budget is 8000 and the number of iterations is 15.

Results



Performance using CS-SVM and FT-th training schemes compared to SVM and FT.

- SVM training scheme outperforms FT
- Method work over the classical imbalance learning techniques

Discussion

- Imbalance needs to be treated at the time of sample selection
- Cost-Sensitive SVM over fixed representation acts as a good alternative to CNN-FT
- Certainty-oriented Minority Class Sampling provides best mitigation to imbalance, while diversity-oriented minority class sampling performs best overall

Perspectives

- Semi-supervised learning- label propagation from source to
- Domain adaptation/ Universality

References

1. Settles, Burr. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences, 2009.
2. Krishnakumar, Anita. "Active learning literature survey." Technical Report. University of California, 2007.
3. Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. IEEE Trans. Circuits Syst. Video Techn.,2017
4. Fábio Perez, Rémi Lebret, Karl Aberer,2019 Weakly Supervised Active Learning with Cluster Annotation
5. Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In Doina Precup and Yee Whye Teh, ICML 2017,
6. Sener and S. Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach. ArXiv e-prints, August 2017.
7. William H. Beluch, Tim Genewein, Andreas Nürnberger, Jan M. Köhler; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9368-937
8. Souza, Vinicius, et al. "Unsupervised active learning techniques for labeling training sets: an experimental evaluation on sequential data." Intelligent Data Analysis 21.5 (2017): 1061-1095.examples. IEEE Trans Pattern Anal Mach Intell 2014
9. Gao, Mingfei, et al. "Consistency-Based Semi-Supervised Active Learning: Towards Minimizing Labeling Cost." *arXiv preprint arXiv:1910.07153* (2019).
10. Siméoni, Oriane, et al. "Rethinking deep active learning: Using unlabeled data at model training." *arXiv preprint arXiv:1911.08177* (2019).
11. Ash, Jordan T., et al. "Deep batch active learning by diverse, uncertain gradient lower bounds." *arXiv preprint arXiv:1906.03671* (2019).

Thank you !

Please visit the Poster number 2743