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

Instance Credibility Inference for Few-Shot Learning

Yikai Wang[1,4] Chengming Xu[1] Chen Liu[1] Li Zhang[2] Yanwei Fu[1,3,4*]

CVPR 2020

Background

What is Few-shot learning ?

It aims to recognize new objects with extremely **limited training data** for each category.

N way K shot:
N classes
K samples each class

Background

An example of 2-way 4-shot task

Background

What are the solutions ?

- Fine-tuning
- Metric learning methods (Matching Network)
- Meta-Learning (MAML)
- Data augmentation and regularization

……

Drawbacks: overfiting, hard to train, not solve it.

Motivation

Fundamental Problem of few-shot learning:

One can hardly estimate the **data distribution** without introducing the **inductive bias**.

Two strategies:

Use unlabeled data

- SSFSL: Semi-supervised few-shot learning
- TFSL: Transductive inference for few-shot learning

Use test data

Motivation

Intuition:

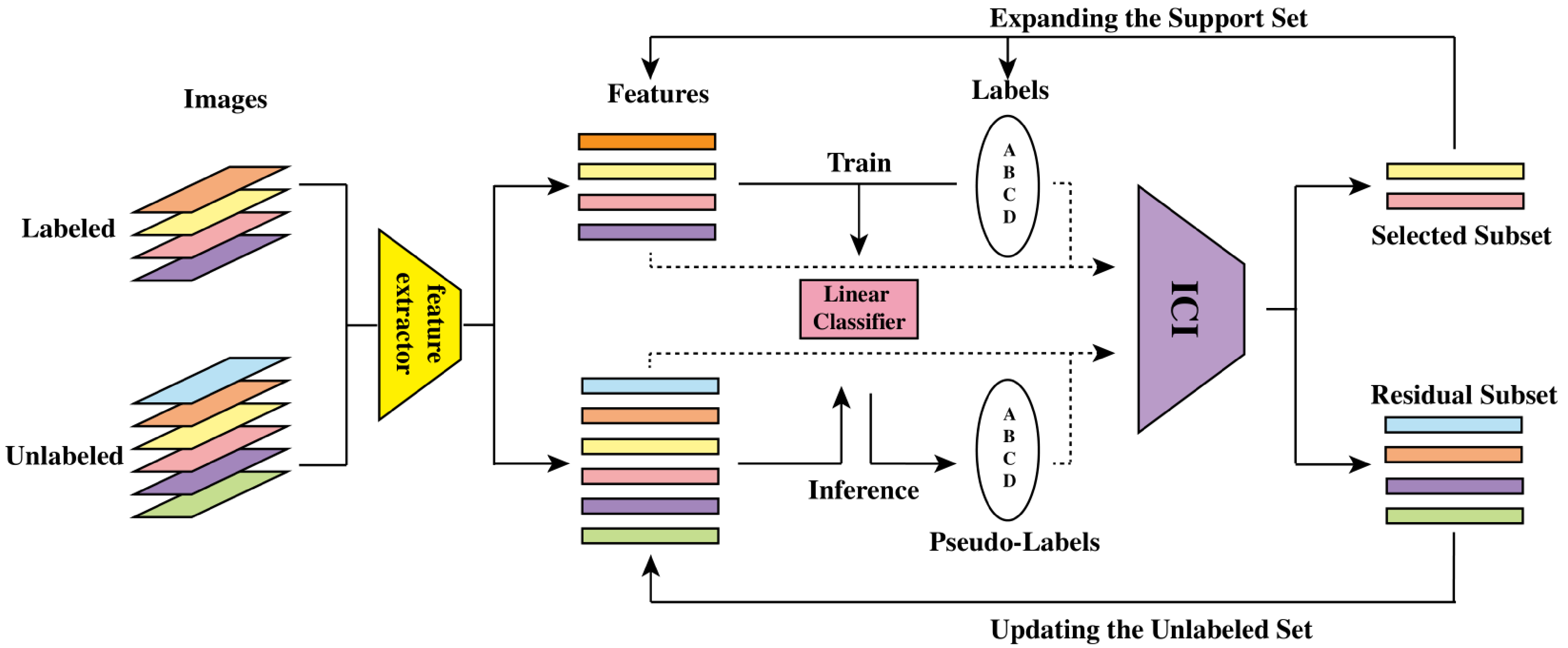
- Few-shot learning model can utilize the data distributions of testing examples.
- Self-taught learning is a way in leveraging the information of unlabeled data.

Problem:

- How to select trustworthy unlabeled data ?
- Are label space (based on the probability of each class given by the classifier) and feature space (select samples most similar to training data) enough ?

Schematic illustration

Instance credibility inference (ICI)



Instance Credibility Inference (ICI)

Train a linear classifier with the labeled few-shot examples and use it to infer the pseudo-labels for the unlabeled data.

Select the most trustworthy pseudo-labeled instances according to their credibility measured by the proposed ICI to augment the training set.

sparsity degree

Classifier thus can be progressively updated and further infer the unlabeled data.

Iterate until converge

Methodology

- A simple linear regression model: $y_i = \mathbf{x}_i^\top \beta + \gamma_i + \epsilon_i$
- Solving the problem of: $\left(\hat{\beta}, \hat{\gamma}\right) = \arg\min_{\beta,\gamma} \|Y - X\beta - \gamma\|_F^2 + \lambda R(\gamma)$
- Rewrite the function as: $L(\beta, \gamma) = \|Y - X\beta - \gamma\|_F^2 + \lambda R(\gamma)$
- Let $\frac{\partial L}{\partial \beta}$ = 0, we have: $\hat{\beta} = \left(X^\top X\right)^\dagger X^\top (Y - \gamma)$
- Take $\hat{\beta}$ into L and solve the problem: $\arg\min_{\gamma \in \mathbb{R}^{n \times N}} \|Y - H(Y - \gamma) - \gamma\|_F^2 + \lambda R(\gamma)$

$$\arg\min_{\gamma \in \mathbb{R}^{n \times N}} \left\|\tilde{Y} - \tilde{X}\gamma\right\|_F^2 + \lambda R(\gamma)$$

Note:
$$\begin{cases} H = X\left(X^\top X\right)^\dagger X^\top \\ \tilde{X} = (I - H) \\ \tilde{Y} = \tilde{X} Y \end{cases} \qquad R(\gamma) = \sum_{i=1}^n \|\gamma_i\|_2$$

Methodology

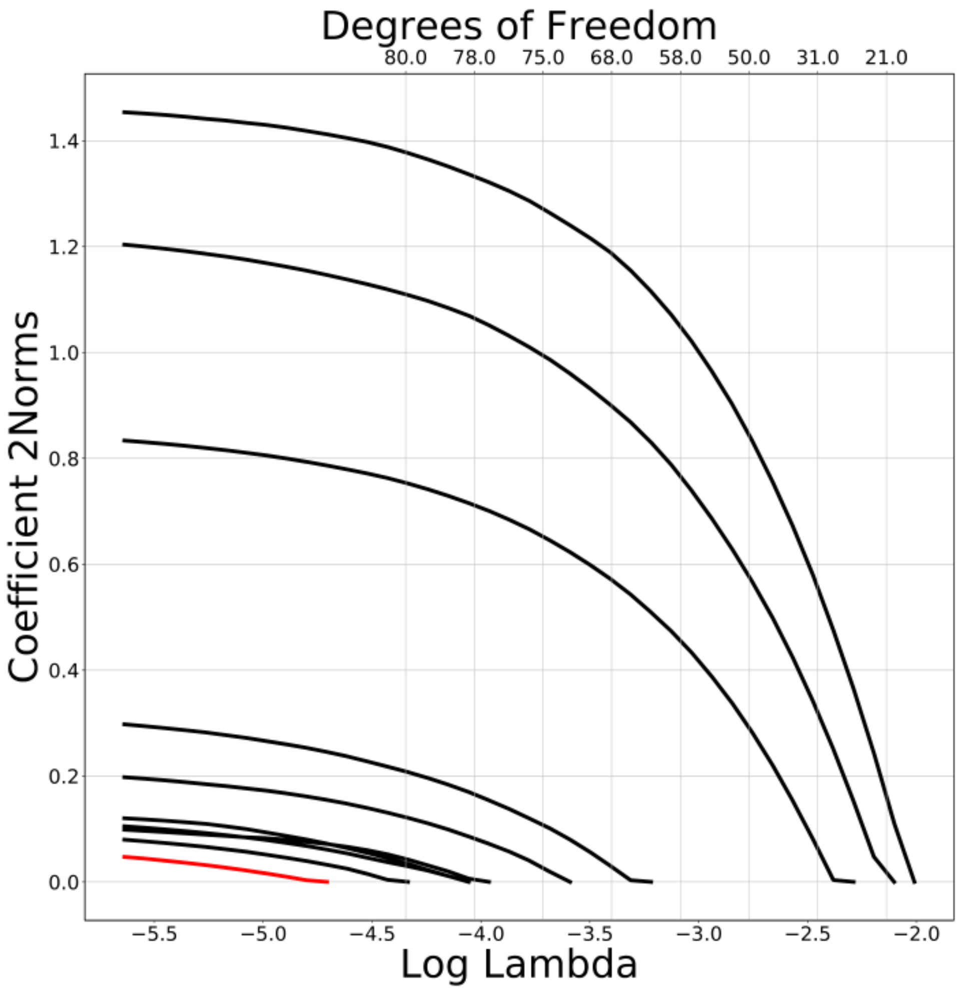
Target: $$\underset{\gamma \in \mathbb{R}^{n \times N}}{\arg\min} \left\| \tilde{Y} - \tilde{X}\gamma \right\|_F^2 + \lambda R(\gamma)$$

A theoretical value λmax guarantee the solution of equation all 0:

$$\lambda_{max} = \max_i \left\| \tilde{X}_{\cdot i}^{\top} \tilde{Y} \right\|_2 / n$$

When **λ** changes from 0 to ∞, the sparsity of **γ** is increased until all of its elements are forced to be vanished

λ: ↑ **Sparsity of γ:** ↑



Algorithm

Algorithm 1 Inference process of our algorithm.

Input:support data$\{(X_i, y_i)\}_{i=1}^{N\times K}$, query data $X_t = \{X_j\}_{j=1}^{M}$, unlabeled data $X_u = \{X_k\}_{k=1}^{U}$

Initialization: support set $(X_s, y_s) = \{(X_i, y_i)\}_{i=1}^{N\times K}$, feature matrix $X_{N\times K+U,d} = [X_s; X_u]$, classifier

Repeat:

Train classifier using (X_s, y_s);

Get pseudo-label y_u for X_u by classifier;

Rank $(X, y) = (X, [y_s; y_u])$ by ICI;

Select a subset (X_{sub}, y_{sub}) into (X_s, y_s);

Until Converged.

Inference:

Train classifier using (X_s, y_s);

Get pseudo-label y_t for X_t by classifier;

Output: inference labels $y_t = \{\hat{y}_j\}_{j=1}^{M}$

Experiments

Nanjing University of Aeronautics and Astronautics

Setting	Model	*mini*ImageNet 1shot	*mini*ImageNet 5shot	*tiered*ImageNet 1shot	*tiered*ImageNet 5shot	CIFAR-FS 1shot	CIFAR-FS 5shot	CUB 1shot	CUB 5shot
In.	Baseline* [7]	51.75	74.27	-	-	-	-	65.51	82.85
In.	Baseline++* [7]	51.87	75.68	-	-	-	-	67.02	83.58
In.	MatchingNet* [53]	52.91[1]	68.88[1]	-	-	-	-	72.36[1]	83.64[1]
In.	ProtoNet* [43]	54.16[1]	73.68[1]	-	-	72.20[3]	83.50[3]	71.88[1]	87.42[1]
In.	MAML* [10]	49.61[1]	65.72[1]	-	-	-	-	69.96[1]	82.70[1]
In.	RelationNet* [46]	52.48[1]	69.83[1]	-	-	-	-	67.59[1]	82.75[1]
In.	adaResNet [31]	56.88	71.94	-	-	-	-	-	-
In.	TapNet [56]	61.65	76.36	63.08	80.26	-	-	-	-
In.	CTM† [25]	64.12	80.51	68.41	84.28	-	-	-	-
In.	MetaOptNet [23]	64.09	80.00	65.81	81.75	72.60	84.30	-	-
Tran.	TPN [28]	59.46	75.65	58.68[4]	74.26[4]	65.89[4]	79.38[4]	-	-
Tran.	TEAM* [34]	60.07	75.90	-	-	70.43	81.25	80.16	87.17
Semi.	MSkM with MTL	62.10[2]	73.60[2]	68.6[2]	81.00[2]	-	-	-	-
Semi.	TPN with MTL	62.70[2]	74.20[2]	72.10[2]	83.30[2]	-	-	-	-
Semi.	MSkM [37]	50.40	64.40	52.40	69.90	-	-	-	-
Semi.	TPN [28]	52.78	66.42	55.70	71.00	-	-	-	-
Semi.	LST [45]	70.10	78.70	77.70	85.20	-	-	-	-
In.	LR	56.06	75.70	69.02	85.37	62.25	80.82	76.16	90.32
In.	SVM	54.46	74.76	67.51	84.67	60.94	79.93	75.84	89.26
Tran.	LR + ICI	66.80	79.26	80.79	87.92	73.97	84.13	88.06	92.53
Tran.	SVM + ICI	65.77	78.94	80.56	87.93	73.16	83.72	87.87	92.38
Semi.	SVM + ICI (15/15)	64.81	78.11	79.72	87.39	72.52	83.23	86.83	91.58
Semi.	SVM + ICI (30/50)	68.24	79.25	83.14	88.58	75.50	84.00	88.94	92.14
Semi.	LR + ICI (15/15)	65.86	78.87	81.10	87.83	73.67	83.85	87.28	92.18
Semi.	LR + ICI (30/50)	69.66	80.11	84.01	89.00	76.51	84.32	89.58	92.48
Semi.	LR + ICI (80/80)	**71.41**	**81.12**	**85.44**	**89.12**	**78.07**	**84.76**	**91.11**	**92.98**

Experiments

Compare ICI with other sample selection strategies

Model	Tran.		Semi.	
	1shot	5shot	1shot	5shot
LR	56.06	75.43	56.06	75.43
+ ra	59.01	76.38	59.46	76.58
+ nn	63.24	77.63	63.10	77.75
+ co	63.29	77.92	63.57	77.71
ICI	**65.32**	**78.30**	**64.60**	**77.96**

Table 2. Compare to baselines on *mini*ImageNet under several settings.

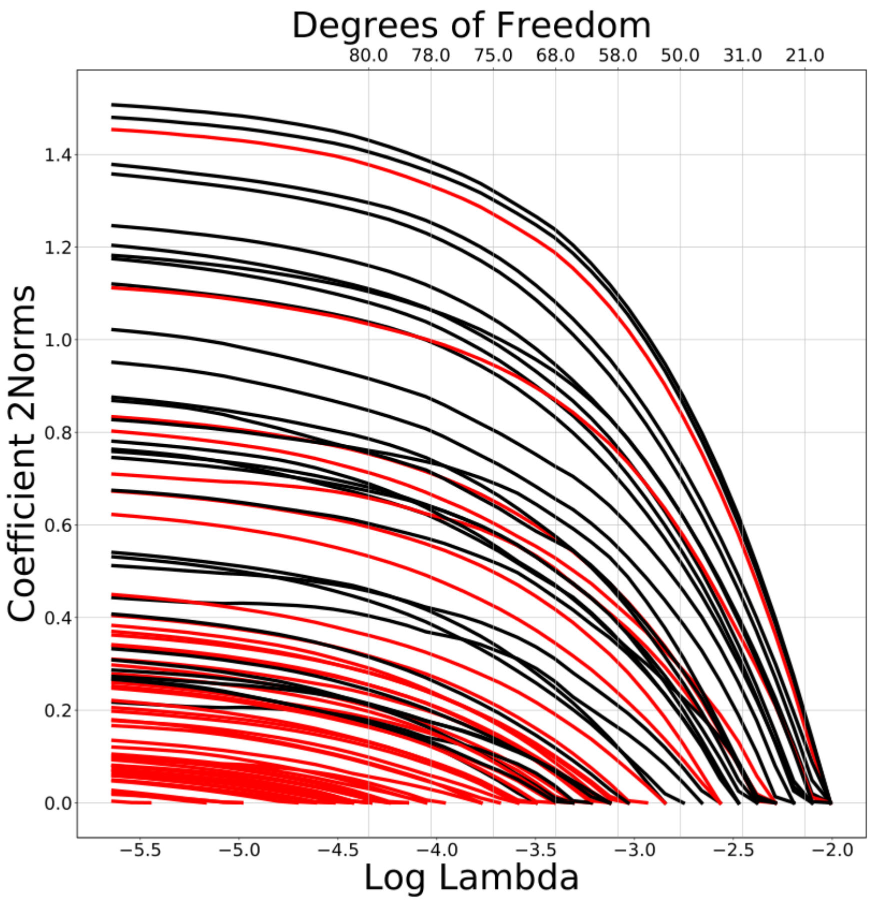


Figure 3. Regularization path of λ. Red lines are correct-predicted instances while black lines are wrong-predicted ones. ICI will choose instances in the lower-left subset.

Experiments

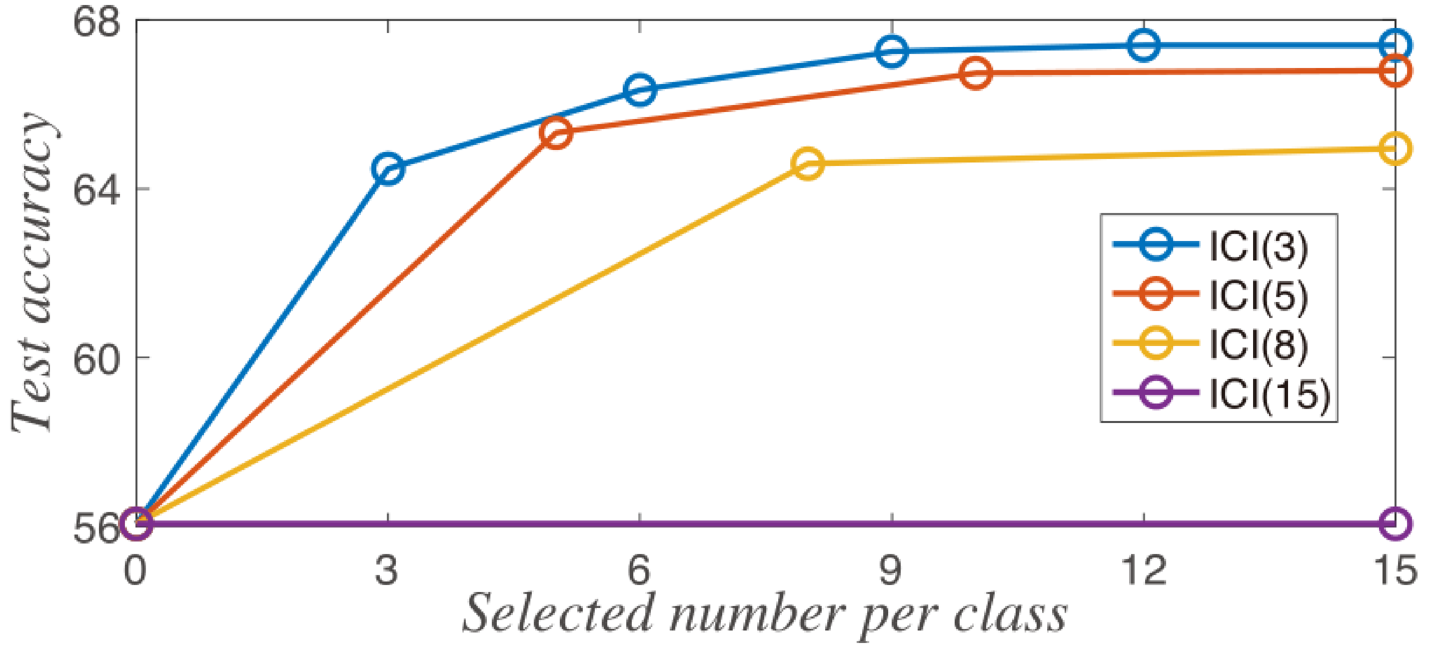


Figure 4. Variation of accuracy as the selected samples increases over 600 episodes on *mini*ImageNet. "ICI (n)": select n samples per class in each iteration.

Conclusion

- Instance Credibility Inference (ICI): exploit the **distribution support** of **unlabeled** instances for few-shot learning

- In order to measure the credibility of each pseudo-labeled instance, we propose to solve a linear regression hypothesis by increasing the **sparsity** of the incidental parameters and rank the pseudo-labeled instance with their **sparsity degree**

THANKS