Supplementary Material for
Few-shot Continual Infomax Learning

Ziqi Gu#, Chunyan Xu#, Jian Yang, Zhen Cui*
PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.
{ziqigu, cyx, csjyang, zhen.cui}@njust.edu.cn

1. Background Knowledge

Entropy: Entropy is a basic concept in information theory that represents a random variable as a measure of uncertainty [17]. For example, $H(X)$ denotes the entropy of a random distribution X.

Transfer Entropy: Transfer entropy is a measure of the amount of information transferred by two stochastic processes. The transfer of stochastic process X to stochastic process Y is achieved by knowing the past of X to reduce the uncertainty of the future of Y, where the information is measured by entropy [2], formally:

$$\mathcal{T}_{X\to Y} = H(Y_t|Y_{t-1:t-L}) - H(Y_t|Y_{t-1:t-L}, X_{t-1:t-L}). \tag{1}$$

Eqn. 1 is equivalently transformed to conditional mutual information, formally:

$$\mathcal{T}_{X\to Y} = I(Y_t; X_{t-1:t-L}|Y_{t-1:t-L}). \tag{2}$$

Mutual Information Estimation: Mutual information(MI) is the reduction of uncertainty in one random variable due to the knowledge of another random variable [1, 7]. Specifically, it is the information obtained from one random variable through another random variable. For two random variables X and Y, the joint probability distribution is $p(X, Y)$. The mutual information between X and Y is given by,

$$I(X;Y) = \int dxdy\ \ p(X,Y)\ \ \log(\frac{p(X,Y)}{p(X)p(Y)}). \tag{3}$$

# Equal Contribution.
* Corresponding Author.

Mutual information is equivalently represented as,

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \\ &= \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)}) - \sum_{x,y} p(x,y)\log p(y) \\ &= \sum_{x,y} p(x)p(y|x)\log(p(y|x)) - \sum_{x,y} p(x,y)\log p(y) \\ &= -\sum_{x} p(x)H(Y|X=x) - \sum_{y}\log p(y)p(y) \\ &= H(Y) - H(Y|X). \end{aligned} \tag{4}$$

where, $H(X)$ is the marginal entropy, $H(X|Y)$ is the conditional entropy, $H(X,Y)$ is the joint entropy of X and Y. Thus, $I(X;Y)$ equals $H(X) - H(X|Y)$ and $H(X) - H(X|Y)$ [1, 7].

Due to the high-dimensional feature vector, it is difficult to accurately compute the mutual information between two variables. Thus, the mutual information of two random variables X and Y can be represented by Kullback-Leibler divergence [5], formally:

$$\begin{aligned} I(X;Y) =& D_{KL}(p(X,Y)||p(X)\otimes p(Y)) \\ =& \mathbb{E}_{p_{(X,Y)}}[\mathcal{F}] - log\mathbb{E}_{p_X\otimes p_Y}[e^{\mathcal{F}}], \end{aligned} \tag{5}$$

where, $p(X,Y)$ is the joint probability distribution of X and Y, all functions $\mathcal{F}$ such that both expectations are finite. Since the mutual information of high-dimensional vectors is difficult to compute, to solve the Eqn. 5, we use neural network to estimate the maximum lower bound [1, 7].

$$\begin{aligned} I(X;Y) \geq& \hat{I}((X;Y), \vartheta^{MI}) \\ =& \mathbb{E}(\mathcal{F}((X;Y)), \vartheta^{MI}) - \log\mathbb{E}[e^{\mathcal{F}((X;Y),\vartheta^{MI})}]. \end{aligned} \tag{6}$$

Here, ϑ^{MI} refers to a neural network to estimate mutual information between X and Y.

Table 1. Few-shot continual classification performance of state-of-the-art methods and our FCIL on the CUB200 [16] dataset. The results with * are obtained from the authors' published code. FCIL outperforms second place by 2.02% in terms of the final accuracy and by 1.41% in terms of the Avg and by 1.35% in terms of the KR.

Methods	Accuracy in each session (%) ↑											KR↑	ΔFinal↑	Avg↑
	1	2	3	4	5	6	7	8	9	10	11			
Ft-CNN	68.68	43.7	25.05	17.72	18.08	16.95	15.1	10.6	8.93	8.93	8.47	12.33	+50.01	22.02
NCM [8]	68.68	57.12	44.21	28.78	26.71	25.66	24.62	21.52	20.12	20.06	19.87	28.93	+38.61	32.49
iCaRL [12]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16	30.80	+37.32	36.67
EEIL [3]	68.68	53.63	47.91	44.2	36.3	27.46	25.93	24.7	23.95	24.13	22.11	32.19	+36.37	36.27
TOPIC [14]	68.68	62.49	54.81	49.99	45.25	41.4	38.35	35.36	32.22	28.31	26.28	38.26	+32.20	43.92
SPPR [21]	68.68	61.85	57.43	52.68	50.19	46.88	44.65	43.07	40.17	39.63	37.33	54.35	+21.15	49.32
Decoupled-DeepEMD [18]	75.35	70.69	66.68	62.34	59.76	56.54	54.61	52.52	50.73	49.20	47.60	63.17	+10.88	58.73
Decoupled-NegCosine [11]	74.96	70.57	66.62	61.32	60.09	56.06	55.03	52.78	51.50	50.08	48.47	64.66	+10.01	58.86
Decoupled-Cosine [15]	75.52	70.95	66.46	61.20	60.86	56.88	55.40	53.49	51.94	50.93	49.31	65.29	+9.17	59.36
CEC [19]	75.8	71.94	68.5	63.5	62.43	58.27	57.73	55.81	54.83	53.52	52.28	68.97	+6.20	61.33
MateFSCIL [4]	75.9	72.41	68.78	64.78	62.96	59.99	58.30	56.85	54.78	53.83	52.64	69.35	+5.84	61.93
FACT* [20]	77.38	73.91	**70.32**	65.91	65.02	61.82	61.29	59.53	57.92	57.63	56.46	72.95	+2.02	64.29
FCIL(Ours)	**78.70**	**75.12**	70.10	**66.26**	**66.51**	**64.01**	**62.69**	**61.00**	**60.36**	**59.45**	**58.48**	**74.30**		**65.70**

Table 2. Few-shot continual classification performance of state-of-the-art methods and our FCIL on the CIFAR100 [10] dataset. The results with * are obtained from the authors' published code. FCIL outperforms second place by 0.7% in terms of the final accuracy and by 0.15% in terms of the Avg and by 0.2% in terms of the KR.

Methods	Accuracy in each session (%) ↑									KR↑	ΔFinal↑	Avg↑
	1	2	3	4	5	6	7	8	9			
Ft-CNN	64.10	36.91	15.37	9.8	6.67	3.8	3.7	3.14	2.65	4.13	+49.37	16.24
iCaRL [12]	64.10	53.28	41.69	34.13	27.93	25.06	20.41	15.48	13.73	21.41	+38.65	32.87
NCM [8]	64.10	53.05	43.96	36.97	31.61	26.73	21.23	16.78	13.54	21.12	+38.48	34.22
EEIL [3]	64.10	53.11	43.71	35.15	28.96	24.98	21.01	17.26	15.85	24.72	+36.17	33.79
TOPIC [14]	64.10	55.88	47.07	45.16	40.11	36.38	33.96	31.55	29.37	45.81	+22.65	42.62
Decoupled-DeepEMD [18]	69.75	65.06	61.20	57.21	53.88	51.40	48.80	46.84	44.41	63.67	+7.61	55.39
Decoupled-NegCosine [11]	74.36	68.23	62.84	59.24	55.32	52.88	50.86	48.98	46.66	62.73	+5.36	57.71
Decoupled-Cosine [15]	74.55	67.43	63.63	59.55	56.11	53.80	51.68	49.67	47.68	63.95	+4.34	58.23
CEC [19]	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14	67.25	+2.61	59.53
MateFSCIL [4]	74.50	70.10	66.84	62.77	59.48	56.52	54.36	52.56	49.97	67.07	+2.05	60.79
C-FSCIL Mode1(d=512) [6]	77.47	72.20	67.53	63.23	59.58	56.67	53.94	51.55	49.36	63.71	+2.66	61.28
C-FSCIL Mode2(d=512) [6]	77.50	**72.45**	67.94	63.80	60.24	57.34	54.61	52.41	50.23	64.81	+1.79	61.84
C-FSCIL Mode3(d=512) [6]	77.47	72.40	67.47	63.25	59.84	56.95	54.42	52.47	50.47	65.14	+1.55	61.64
FACT* [20]	**78.44**	72.33	68.23	63.90	60.58	**58.20**	55.96	53.59	51.32	65.43	+0.70	62.51
FCIL(Ours)	77.12	72.42	**68.31**	**64.47**	**61.18**	58.17	**56.06**	**54.19**	**52.02**	**67.45**		**62.66**

2. Datasets and other results

2.1. Datasets

CIFAR100: CIFAR100 [9] contains 100 classes with a total of 60,000 RGB images with the size 32×32, and each class contains 500 training images and 100 test images.

CUB200: CaltechUCSD Birds-200-2011(CUB200) [16] is a fine-grained classification dataset, which contains 11,788 images in 200 classes, each image size is 224×224.

miniImagenet: miniImageNet [13] is a subset of ImageNet, which contains 100 classes with 60,000 images, and each image size is 84×84.

2.2. Results of other datasets

In the main paper, we report the detailed performance of the miniImageNet. We report the performance of the CUB200 [16] and CIFAR100 [10] in Table 1 and Table 2. We can infer that our proposed FCIL has better final accuracy, Avg and KR, indicating FCIL better than state-of-the-art methods.

Algorithm 1 Few-shot Continual Infomax Learning (FCIL)

Require: the training sets $\{(X_t, Y_t)|t = 1, \cdots, T\}$, the number of previous sessions K.
Ensure: the final model Θ and ϑ.

```
while t = 1, 2, ..., T do
    if t = 1 then
        Optimize the base network Θ^base and the MI network ϑ^MI on the training set (X_1, Y_1);
        Construct base class structure S(A^t, R^t);
    else
        # for the t-th session data
        Learn new-class model Θ^t_fc by feature embedding infomax L_FEI;
        Update class structure S(A^t, R^t);
        Update the classifier Θ_fc by performing continual structure infomax L_CSI;
    end if
end while
```

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018. 1

[2] Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T Lizier. Transfer entropy. In *An introduction to transfer entropy*, pages 65–95. 2016. 1

[3] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision*, pages 233–248, 2018. 2

[4] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14166–14175, 2022. 2

[5] Monroe D Donsker and SR Srinivasa Varadhan. On a variational formula for the principal eigenvalue for operators with maximum principle. *National Academy of Sciences*, 72(3):780–783, 1975. 1

[6] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9057–9067, 2022. 2

[7] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 1

[8] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019. 2

[9] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 2

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of The ACM*, 2012. 2

[11] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Proceedings of the European Conference on Computer Vision*, pages 438–455. Springer, 2020. 2

[12] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[14] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12183–12192, 2020. 2

[15] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 2

[16] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[17] Alfred Wehrl. General properties of entropy. *Reviews of Modern Physics*, 50(2):221, 1978. 1

[18] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE Conference on Computer Vision and pattern recognition*, pages 12203–12213, 2020. 2

[19] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12455–12464, 2021. 2

[20] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9046–9056, 2022. 2

[21] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6801–6810, 2021. 2