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

Abstract

Few-shot continual learning is the ability to continually train a neural network from a sequential stream of few-shot data. In this paper, we propose a Few-shot Continual Infomax Learning (FCIL) framework that makes a deep model to continually/incrementally learn new concepts from few labeled samples, relieving the catastrophic forgetting of past knowledge. Specifically, inspired by the theoretical definition of transfer entropy, we introduce a feature embedding infomax to effectively perform the few-shot learning, which can transfer the strong encoding capability of the base network to learn the feature embedding of these novel classes by maximizing the mutual information of different-level feature distributions. Further, considering that the learned knowledge in the human brain is a generalization of actual information and exists in a certain relational structure, we perform continual structure infomax learning to relieve the catastrophic forgetting problem in the continual learning process. The information structure of this learned knowledge can be preserved through maximizing the mutual information across these continual-changing relations of inter-classes. Comprehensive evaluations on CIFAR100, miniImageNet, and CUB200 datasets demonstrate the superiority of our FCIL when compared against state-of-the-art methods on the few-shot continual learning task.

1. Introduction

Recent deep learning technologies [23, 30, 13] have made great progress in many computer vision tasks. The success of deep neural networks is achieved by employing a large number of labeled training data. However, it is difficult to collect large-scale supervised data in advance, while we would encounter a sequential data stream with unknown new classes in many realistic situations. This would require a neural network with continual learning ability, especially when some new classes with very few labeled samples often appear. Therefore, this work focuses on dealing with the few-shot continual/incremental learning task [28, 26]. They are two critical challenges in a few-shot continual learning task: i) how to learn new knowledge with few-shot instances, and ii) how to avoid catastrophic forgetting of the preceding learned knowledge.

Recently, numerous previous works [31] have been devoted to few-shot continual learning from various perspectives. In [31, 35, 8], the topological structure of the knowledge space formed by different classes has been considered to perform few-shot continual learning by using a neural gas network or graph model. Cheraghian et al. [8] have proposed a semantic-aware knowledge distillation method to solve few-shot class-incremental learning by making use of word embeddings. Several meta-learning based methods [7, 18] have been also proposed to enable the model to preserve old knowledge and adapt to the new classes for continual learning. By assigning virtual prototypes to squeeze the embedding of known classes and reserve for new ones, the forward compatible training method [37] has efficiently incorporated new classes with forward compatibility and meanwhile resists forgetting old ones. Similarly, the mixture of subspaces and synthesized features [6] have been used to alleviate the forgetting and over-fitting problem in the few-shot continual learning process. In [38], the self-promoted prototype learning scheme has been proposed to explicitly learn the feature representation under the few-shot learning situation and thus facilitated subsequent incremental tasks. The self-supervised strategy has been also used to enhance the feature extraction ability of the model by adding self-supervised loss function assistance during the training process [20, 26]. Moreover, to minimize over-fitting and the catastrophic forgetting problem, Mazumder et al. [26] have selected very few unimportant model parameters to perform few-shot learning on new classes. In [11], mutual information (MI) maximization has been first used as a solution method to deal with the catas-

[#] Equal Contribution.
[*] Corresponding Author.

trophic forgetting problem in the online continual learning task, but it cannot well consider how to perform continual infomax learning from the perspective of information entropy, especially in the few-shot regime.

In this work, we propose a novel Few-shot Continual Infomax Learning (FCIL) framework that makes a deep model to continually learn from a stream of few-shot labeled data. In general, the continual learning model would be trained with a large amount of labeled data in the initial learning stage, while few-shot samples of some unknown classes would be encountered in the continual learning process. Here we attempt to address the few-shot continual learning task from two aspects. First, inspired by the theoretical definition of transfer entropy, we attempt to transfer the strong encoding capability of the base network to promote few-shot continual learning. Specifically, we propose a feature embedding infomax learning to new concepts from a few labeled samples through maximizing the mutual information between different level feature distributions, where the convolutional representations and the feature embedding of new classes are encoded with these fixed convolution layers of the base network and newly increased parameters of fully-connected layer, respectively. Second, considering that the learned knowledge in the human brain is a generalization or abstraction of the actual information learned from these seen samples [10], we wish the continual learning model with a stable information structure that can be updated incrementally. Thus we propose a continual structure infomax learning mechanism to alleviate the catastrophic forgetting problem during the continual learning process. The structure relation of this learned information can be preserved through maximizing the mutual information across these continual-changing structure relations of inter-classes.

In summary, our primary contributions can be summarized as follows: i) We propose a Few-shot Continual Infomax Learning (FCIL) framework that makes a deep model to incrementally learn new concepts from few labeled samples, relieving the catastrophic forgetting problem of previously learned ones. ii) Two specially-designed infomax learning mechanisms are proposed to address the few-shot continual learning problem with the help of mutual information maximization, including feature embedding infomax and continual structure infomax. iii) We validate the effectiveness of our proposed FCIL on three benchmarks (including CIFAR100, CUB200, and miniImageNet), and also demonstrate the superiority of our proposed FCIL when compared with existing state-of-the-art methods.

2. Related work

Few-shot continual learning: Few-shot continual learning, which has recently attracted growing attention, aims to enable a model to perform continual/incremental learning with a stream of few-shot labeled data. A number of related works have been proposed to address the few-shot continual learning task. For example, Tao et al. [31] employed neural gas networks to perform incrementally learning for a series of new classes with few-shot data, with the goal of avoiding forgetting previously learned classes. Zhang et al. [35] proposed an evolving classifier based on graph attention networks that propagated information between classifiers by adding graph models. Also, pseudo-incremental training was proposed to optimize the graph module. The forward compatible training (FACT) [37] was proposed to effectively incorporate novel classes into forward compatible by generating virtual classes, where knowledge of old classes can be also maintained to reduce the catastrophic forgetting problem. A bi-level meta-learning-based optimization [7] was proposed to directly optimize the model towards forgetting alleviation of learned knowledge and adaptation of the new classes. In [20], a self-supervised stochastic classifier was proposed to deal with the few-shot class incremental learning task with the help of a self-supervision mechanism. Zhu et al. [38] proposed an incremental prototype learning scheme to perform continual learning, which was used to constrain the prototypes of new classes through the dynamic model and then achieved a reduction in catastrophic forgetting of knowledge. In order to offer a trade-off between accuracy and compute-memory cost of learning novel classes, the hyper-dimensional embedding was used to continually learn many more classes than the fixed dimensions in the feature space [14]. Further, few-shot lifelong learning was proposed to perform continual learning with only updating very few parameters of the model [26] or semantic-aware knowledge distillation [5].

Mutual information: Mutual Information as a quantity is used to measure the degree of interdependence between two random variables [21]. As it is difficult to exactly compute the mutual information between second-dimensional random variables, mutual information neural estimator (MINE) [2] can be achieved by employing the dual representations of the Kullback-Leibler divergence (KL) [9]. Subsequently, the Jensen-Shannon divergence (JSD) [27] and noise-contrastive estimation (NCE) [12] were used to estimate the mutual information with a learned network [16], and then the mutual information maximization method was used to improve the deep representation in an unsupervised learning way. The second-order deep multiplex infomax method [19] further extended the mutual information estimation to a three-variable calculation and was applied across multiple networks. The deep mutual information maximin method [25] was used to address the cross-modal clustering task by maximally preserving the shared information of multiple modalities and eliminating the superfluous information of individual modalities. In [1], mutual information networks are applied to unsupervised

classification tasks with excellent results. In the online continual learning task, Guo et al. [11] first used mutual information estimation to learn more robust features and preserve the past model learned from the previous task. Different from these previous methods, we perform continual infomax learning from the perspective of information entropy in the few-shot setting.

3. The Proposed Method

3.1. Problem Definition

Few-shot continual learning aims to train a model for new tasks sequentially from a stream of few-shot labeled data, relieving the forgetting of knowledge learned from the old tasks, where the data in the old tasks are not available anymore during learning a new set of few-shot tasks. Formally, we define X, Y, and Z as the training set, label set, and test set, respectively. Our task is to train the model with a continuous stream of labeled training sets $X_1, X_2, \ldots, X_T$, where X_t denotes the t-th training set, Y_t is the corresponding label set, and T refers to the number of continual learning sessions. In the continual learning process, each training set is with no repetition of class labels, i.e., $\forall i, j$ and $i \neq j$, $Y_i \cap Y_j = \varnothing$. Usually, when $t = 1$, the training set X_1 is with large-scale samples for training the base network; when $t > 1$, X_t refers to the few-shot training set of new classes in the incremental learning stage (i.e., $|X_1| \gg |X_t|, t = 2, 3, \ldots, T$). For example, under the 5-way 10-shot setting, each incremental/continual process (i.e., $t > 1$) contains five new classes, each of which has only ten training samples. The test set Z is used to evaluate the classification performance at each stage t, and its corresponding classes may be from all training label sets, i.e., $\{Y_1 \cup Y_2 \cdots \cup Y_T\}$.

In the few-shot continual learning task, there are two critical problems we need to consider: i) New classes are with a small amount of training data, which makes it hard to learn the feature embeddings for these new classes. ii) Catastrophic forgetting should be well relieved to preserve this learned knowledge in the continual learning process. In this work, we attempt to address them from two aspects. First, we start from the transfer entropy perspective [3] to obtain new knowledge from few-shot samples, where the strong encoding capability of the base network can be transferred to learn feature embeddings of these new classes. Second, considering that the learned knowledge in the human brain is not isolated, but exists in a certain relational structure [10], we wish to preserve the structure information of this learned knowledge to relieve the catastrophic forgetting problem during the continual learning process. With the help of the recently-developed mutual information estimation method [2, 16, 36], we achieve these above goals by performing continual infomax learning.

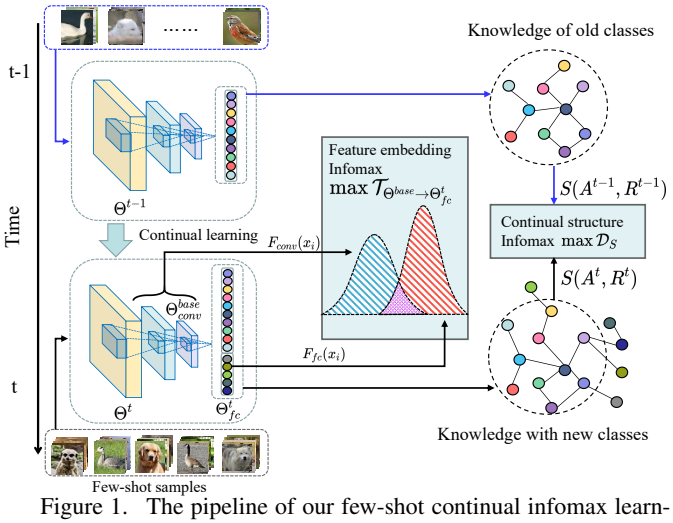


Figure 1. The pipeline of our few-shot continual infomax learning, where t refers to the t-th session in the continual learning process ($t \geqslant 1$). To transfer the strong encoding capability of the base network into the subsequent continual learning sessions, we first propose the feature embedding infomax to learn new knowledge with few-shot labeled data in the t-th session. Here we maximize the mutual information between multi-level convolutional representations $F_{conv}(x_i)$ and full-connected feature embedding $F_{fc}(x_i)$ of each input sample, which are encoded with these fixed convolution layers of the base network Θ^{base}_{conv} and updated fully-connected layer Θ^{t}_{fc}, respectively. To address the catastrophic forgetting problem in the continual learning process, we further perform the continual structure infomax learning by maximizing the mutual information between current and previous inter-class knowledge structures (i.e., $S(A^t, R^t)$ and $S(A^{t-1}, R^{t-1})$).

3.2. FCIL Paradigm

The proposed few-shot continual infomax learning (FCIL) framework is shown in Fig. 1. Generally, the base network Θ^{base}, which is a large-scale set of training samples X_1, can optimize with the classic cross-entropy loss. For further improving the deep representation capability of the base network, we also employ self-supervised mutual information mechanism. Here the mutual information estimated network ϑ^{MI} can be well optimized in the initial stage (i.e., $t = 1$), which contributes to the subsequent continual learning sessions. In the next continual learning sessions (i.e., $t > 1$), a stream of few-shot labeled data with new classes would be gradually fed into the network, and we would perform few-shot continual infomax learning for capturing new knowledge, without forgetting these learned ones. To improve the feature embedding capability of these new classes, we attempt to transfer the strong representation capability of the base network Θ^{base} into the next continual learning sessions. Specifically, we introduce the feature embedding infomax learning to optimize these fully-connected parameters Θ^{t}_{fc} of new classes. Given an input image x_i from the few-shot set D_t, we use the convolutional net optimized in the base stage to extract multi-level convolutional

features $F_{conv}(x_i)$, and get the feature embeddings from new full-connected layer $F_{fc}(x_i)$. We maximize the mutual information between the convolutional and full-connected feature distributions for each input sample (i.e., $F_{conv}(x_i)$ and $F_{fc}(x_i)$), and thus the feature embedding of the novel classes can well inherit the powerful feature representation capability of the base network (i.e., $\max \mathcal{T}_{\Theta^{base} \to \Theta^t_{fc}}$). Further, to alleviate the catastrophic forgetting problem of this learned knowledge during the subsequent continual learning sessions, we introduce a continual structure infomax learning (i.e., $\max \mathcal{D}_S$) to maintain/preserve the relations of intra-classes across one or more learning sessions. At each learning session t, we can obtain the first-order attributes and the second-order structure information of this learned knowledge, which refer to the class-wise feature embeddings and the relations among different classes, respectively. We then maximize the mutual information between the current information structure $S(A^t, R^t)$ and this previous information $S(A^{t'}, R^{t'})$ ($t' = t-1, t-2, \ldots, t-K$ and $t - K \geqslant 1$) of these old classes. With the help of feature embedding infomax and continual structure infomax, the deep model can be well worked in the few-shot continual learning process. See the supplementary materia for more details.

3.2.1 Feature Embedding Infomax

When the base network Θ^{base} faces unknown new classes, it is difficult to learn new concepts with only few-shot labeled samples. Inspired by the classic transfer entropy method which is a non-parametric measure of directed, asymmetric information transfer between two processes [3], we attempt to migrate the powerful feature representation capability of the base network to capture the feature embedding of the novel classes. Given few-shot training samples of new classes in the t-continual learning stage (i.e., $t > 1$), the transfer entropy from the base network Θ^{base} to the feature embedding layer of novel classes Θ^t_{fc} (i.e., fully-connected model parameters) can be formulated as:

$$\mathcal{T}_{\Theta^{base} \to \Theta^t_{fc}} = \sum_{i=1}^{|X_t|} (H(\mathbf{Z}_{fc}) - H(\mathbf{Z}_{fc}|\mathbf{Z}_{conv})), \quad (1)$$

where $X_t = \{x_1, x_2, \ldots, x_{|X_t|}\}$ is the training set of few-shot data in the t-th continual learning session; $F_{conv}(x_i)$ is the convolution representations of the training sample x_i with the base encoding network Θ^{base}. $F_{fc}(x_i)$ is the feature embedding of x_i with the fully-connected parameters Θ^t_{fc} of the t-th continual learning stage. Let $\mathbf{Z}_{conv} = F_{conv}(x_i)$ and $\mathbf{Z}_{fc} = F_{fc}(x_i)$, $H(\mathbf{Z}_{fc})$ refers to the Shannon entropy information of fully-connected feature of x_i, and the conditional entropy $H(\mathbf{Z}_{fc}|\mathbf{Z}_{conv})$ is the entropy of $\mathbf{Z}_{conv}$ conditioned on $\mathbf{Z}_{fc}$. To facilitate the calculation of Eqn. (1), the transfer entropy $\mathcal{T}_{\Theta^{base} \to \Theta^t_{fc}}$ can be further expressed as the conditional mutual information of different level features with the corresponding network parameters in the condition:

$$\begin{aligned} \mathcal{T}_{\Theta^{base} \to \Theta^t_{fc}} &= \sum_{i=1}^{|X_t|} \sum_{l=1}^{L} I((\mathbf{Z}_{fc}; \mathbf{Z}^l_{conv})), \\ &\geq \sum_{i=1}^{|X_t|} \sum_{l=1}^{L} \hat{I}((\mathbf{Z}_{fc}; \mathbf{Z}^l_{conv}), \vartheta^{MI}), \end{aligned} \quad (2)$$

here, we directly maximize the mutual information between multi-level convolutional features and fully-connected feature embedding, which can better transfer the strong encoding capability of the base network Θ^{base} to the fully-connected layer Θ^t_{fc}. With the support of the Kullback-Leibler divergence [9] and the learned mutual information estimation network ϑ^{MI}, the transfer entropy process can be approximately formulated as:

$$\begin{aligned} \mathcal{L}_{\mathcal{T}} = &\sum_{i=1}^{|X_t|} \sum_{l=1}^{L} \mathbb{E}(\mathcal{F}((\mathbf{Z}_{fc}; \mathbf{Z}^l_{conv}), \vartheta^{MI})) \\ &- \log \mathbb{E}(e^{\mathcal{F}((\mathbf{Z}_{fc}; \mathbf{Z}^l_{conv}), \vartheta^{MI})}). \end{aligned} \quad (3)$$

where $\mathcal{F}(\cdot)$ is a discriminant function between features of different layers. Here this transfer entropy process can help to capture the feature embeddings of these few-shot training samples in the continual learning stages, thus we name it as "*feature embedding infomax*" learning process. The final learning objective to optimize the t-th continual fully-connected layer Θ^t_{fc} is defined as: $\mathcal{L}_{FEI} = \mathcal{L}_{CE} - \alpha \mathcal{L}_{\mathcal{T}}$, where $\mathcal{L}_{CE}$ is the classic cross-entropy loss and α is a hyper-parameter.

3.2.2 Continual Structure Infomax

Another critical problem for a continual learning model is exiting the catastrophic forgetting this learned knowledge when discarding old samples. An obvious solution is that continual learning must have a strategy to preserve the information related to these old training samples. Usually, the information preserved in a continual learning system is a generalization or abstraction of the actual information embedded in the learned samples. The knowledge we learn in the human brain [10] does not exist in isolation; there is some kind of relational structure between different pieces of knowledge that maintains the connection between knowledge. In order to possess an information structure that would be updated incrementally along the continual learning process, we can preserve useful information from learned samples as much as possible so that the subsequent continual learning process can be guided by this information. Therefore, we propose a continual structure infomax learning to conserve both first-order and second-order information in the continual learning process, which can be

related to class-wise individual attributions and intra-class relations, respectively. In the t-th continual learning session, the continual structure infomax can be performed under the guidance of these previous information structures:

$$\mathcal{D}_S = \sum_{k=1}^{K} \underbrace{H(A_t) - H(A_t|A_{t-k})}_{\text{first-order}} + \underbrace{H(R_t) - H(R_t|R_{t-k})}_{\text{second-order}} \tag{4}$$

where K is the number of previous sessions with $t-K \geq 1$. A_t and A_{t-k} refer to the feature embeddings of all learned classes c_t, e.g., A_t= $[e_1, e_2, \ldots, e_c, \ldots, e_{c_t}]$. Due to the fact that the mean value of each class is approximate to the parameter weights of the last fully-connected layer, e_c is achieved by the weight parameters of the last classification layer corresponding to the c-th class. Their corresponding information entropy A_t and A_{t-k} are represented as $H(A_t)$ and $H(A_t|A_{t-k})$. $R_t \in \mathbb{R}^{c_t \times c_t}$ and $R_{t-k} \in \mathbb{R}^{c_{t-k} \times c_{t-k}}$ refer to the inter-class relation matrices of all learned classes in the t-th and $(t-k)$-th stages. $H(R_t)$ and $H(R_t|R_{t-k})$ represent the second-order information entropy concerning the relation among different classes. Through maximizing the first-order and second-order mutual information between different learning stages, the network parameters of the last fully-connected layer Θ_{fc} can be incrementally updated along the continual learning process, formally,

$$\begin{aligned}\mathcal{D}_S &= \sum_{k=1}^{K} \underbrace{I(A_t; A_{t-k})}_{\text{first-order}} + \underbrace{I(R_t; R_{t-k})}_{\text{second-order}}, \\ &\geq \sum_{k=1}^{K} \underbrace{\hat{I}((A_t; A_{t-k}), \vartheta^{MI})}_{\text{first-order}} + \underbrace{\hat{I}((R_t; R_{t-k}), \vartheta^{MI})}_{\text{second-order}},\end{aligned} \tag{5}$$

Similar to the feature embedding infomax, we also use mutual information to measure the amount of information shared by different sessions, and this continual structure infomax process can be further formulated as:

$$\begin{aligned}\mathcal{L}_S = \sum_{k=1}^{K} &\mathbb{E}(\mathcal{F}((A_t; A_{t-k}), \vartheta^{MI})) - \log \mathbb{E}(e^{\mathcal{F}((A_t; A_{t-k}), \vartheta^{MI})}) \\ &+\mathbb{E}(\mathcal{F}((R_t; R_{t-k}), \vartheta^{MI})) - \log \mathbb{E}(e^{\mathcal{F}((R_t; R_{t-k}), \vartheta^{MI})}),\end{aligned} \tag{6}$$

where k refers to the k-th session with $1 \leqslant k \leqslant K$. We employ both $\mathcal{L}_S$ and $\mathcal{L}_{CE}$ to optimize the fully-connected parameters Θ_{fc}^t to adapt new classes without forgetting these learned ones. In the structure continual infomax process, the corresponding loss is defined as $\mathcal{L}_{CSI} = \mathcal{L}_{CE} - \gamma \mathcal{L}_S$, where γ is a balance factor.

4. Experiments

4.1. Experimental Setup

Datasets: We evaluate our FCIL on three public datasets, including CIFAR100 [22], miniImageNet [29], and CUB200 [33]. For CIFAR100 [22] and miniImageNet [29], we use all the training data of the 60 base classes to train a base network, and the 40 new classes are used to perform eight 5-way 5-shot continual learning tasks. We divide the 200 classes of CUB200 [33] into 100 base classes and 100 new classes. The 100 new classes are used to perform ten 10-way 5-shot continual learning tasks. We follow the same split setting of FSCIL [31] as other methods in the three datasets for a fair comparison. See the supplementary materia for more details.

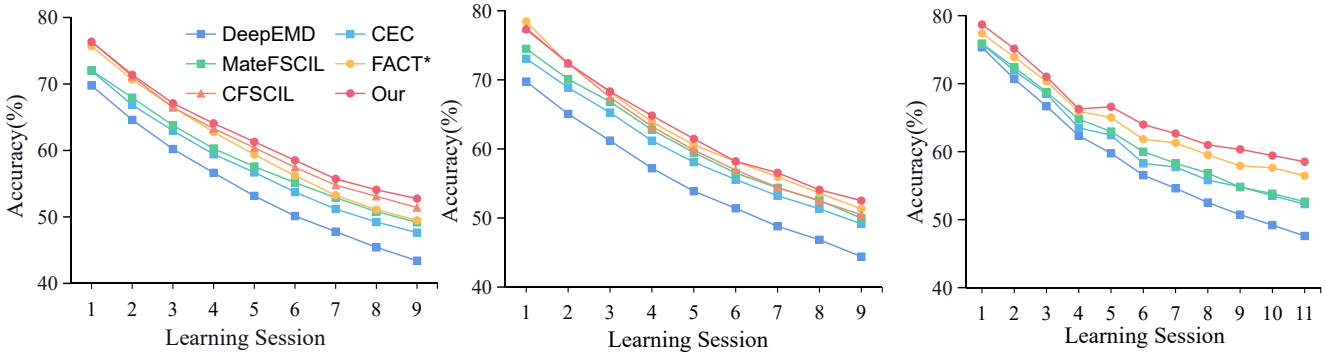
Implementation Details: Following FSCIL [31], we use ResNet20 as the backbone on CIFAR100 [22] and ResNet18 is used as the backbone on miniImageNet [29] and CUB200 [33]. The network ϑ^{MI} consists of a convolution branch and a linear branch, where the former consists of three convolution layers whose kernel sizes are 1, 5, and 7, and the latter is one linear layer. In the CSI module, the convolution branch and the linear branch separately take $F_{conv}^l(x_i)$ and $F_{fc}(x_i)$ as inputs, and their outputs are used to compute $\mathcal{L}_{\mathcal{T}}$. In the CSI module, the linear layer of ϑ^{MI} separately take A_t and A_{t-k} as inputs, and we use the outputs to compute $\mathcal{L}_s$. In the base learning stage, we train the base network and the mutual information network with the first training set. The parameters of the convolutional layers Θ^{base} and the mutual information network ϑ^{MI} are frozen in the continual learning sessions to perform mutual information estimation in both the proposed feature embedding infomax. For all datasets, we set α and γ as 0.015 and 0.03, respectively. Random cropping, random scaling, and random horizontal flipping are used for training data. We use to evaluate our proposed method from different perspectives. "Acc_t" represents to the Top-1 accuracy in the t-th continual session. "Avg" represents for an average score, i.e., $\sum_{t=1}^{T} \text{Acc}_t/T$. "$\Delta$Final" represents the difference value of Acc_t between our method and the compared method in the final state. "KR" is the knowledge retention rate, i.e., $\text{Acc}_T/\text{Acc}_1$. All the experiments are conducted with the Pytorch framework on one Geforce 2080Ti GPU.

4.2. Comparison with state-of-the-art methods

We first compare our proposed FCIL with other state-of-the-art methods on the challenging miniImageNet dataset [29]. The detailed results have been listed in Table 1, and the accuracy changing curves with growing continual sessions are plotted in Fig. 2 (a). Our proposed FCIL outperforms all other compared methods by a great margin and sets new state-of-the-art. In particular, our FCIL achieves the best score among all the methods. Considering that in the first continual session, very little data are available and it requires borrowing the learning capability of the base stage into this session to promote the classifier accuracy. In addition, as to the overall performance in the final continual session, our FCIL also surpasses the second best method C-

Table 1. Few-shot continual classification performance of state-of-the-art methods and our FCIL on the miniImageNet dataset [29]. The results with * are obtained from the authors' published code. See the supplementary materia for results of other datasets.

Methods	Accuracy in each session (%) ↑									KR↑	ΔFinal↑	Avg↑
	1	2	3	4	5	6	7	8	9			
Ft-CNN	61.31	27.22	16.37	6.08	2.54	1.56	1.93	2.60	1.40	2.28	+51.36	13.44
iCaRL [28]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	28.07	+35.55	33.29
EEIL [4]	61.31	46.58	44.00	37.29	33.14	27.12	24.10	21.57	19.58	31.93	+33.18	34.97
TOPIC [31]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	39.83	+28.34	39.65
NCM [17]	61.31	47.80	39.31	31.91	25.68	21.35	18.67	17.24	14.17	23.11	+38.59	30.83
Decoupled-NegCosine [24]	71.68	66.64	62.57	58.82	55.91	52.88	49.41	47.50	45.81	63.90	+6.95	56.80
Decoupled-Cosine [32]	70.37	65.45	61.41	58.00	54.81	51.89	49.10	47.27	45.63	64.84	+7.13	55.99
Decoupled-DeepEMD [34]	69.77	64.59	60.21	56.63	53.16	50.13	47.79	45.42	43.41	62.21	+9.35	54.57
MateFSCIL [7]	72.04	67.94	63.77	60.29	57.58	55.16	52.90	50.79	49.19	68.28	+3.57	58.85
C-FSCIL Mode1 (d=512) [15]	76.37	70.94	66.36	62.64	59.31	56.02	53.14	51.04	48.87	63.99	+3.89	60.52
C-FSCIL Mode2 (d=512) [15]	**76.45**	71.23	66.71	63.01	60.09	56.73	53.94	52.01	50.08	65.50	+2.68	61.14
C-FSCIL Mode3 (d=512) [15]	76.40	71.14	66.46	63.29	60.42	57.46	54.78	53.11	51.41	67.29	+1.35	61.61
FACT* [37]	75.68	70.65	66.53	62.75	59.39	56.19	53.26	51.10	49.48	65.38	+3.28	60.56
FACT* + FCIL	76.21	70.92	66.69	63.23	60.73	57.88	54.08	52.17	50.34	66.05	+2.42	61.36
CEC [35]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	66.15	+5.13	57.75
CEC + FCIL	72.87	68.23	64.46	60.64	57.71	55.12	52.79	50.65	48.62	66.72	+4.14	59.01
FCIL (Ours)	76.34	**71.40**	**67.10**	**64.08**	**61.30**	**58.51**	**55.72**	**54.08**	**52.76**	**69.11**	-	**62.37**



(a) miniImageNet (b) CIFAR100 (c) CUB200

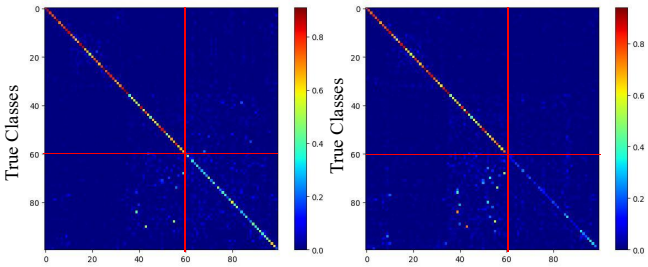
Figure 2. The accuracy changing curves with growing continual sessions of different methods on the three datasets.

FSCIL Mode3 (d=512) [15] by a non-trivial gap of 1.35%. It shows that our FCIL could keep great learning capability when performing the continual learning process, and we owe the superiority to the strong capability borrowed from the base network by the proposed feature embedding infomax module. Further, our method also achieves the best performance on Avg, outperforming the compared methods by at least 0.76%. It indicates that our FCIL could always obtain better performance during all the continual sessions, further demonstrating the strong learning capability borrowed from the base network. Overall, the superior performances of our FCIL in terms of Avg and ΔFinal sufficiently verify the effectiveness of our proposed feature embedding infomax module on enhancing the learning capability of new classes. Besides, we integrate our proposed FCIL mechanisms into the FACT [37] and CEC [35] frameworks (i.e., FACT+FCIL and CEC+FCIL), and the compared results show that we can lead to significant improvements of 0.99% and 0.86% on the ΔFinal metric. It further verifies the superiority of our FCIL framework in learning new knowledge and alleviating catastrophic forgetting.

Our FCIL achieves the best performance on the KR metric, with a promotion of at least 0.83%. Since KR represents the accuracy ratio of the final continual session and the base stage, a larger KR value means that less learned knowledge is forgotten. Compared with previous methods such as CEC [35], FACT [37], C-FSCIL [15] which also freeze the base classifier as our FCIL to avoid catastrophic forgetting, our FCIL still outperforms them by a considerable margin. It means that the relationship preserving capability brought by the proposed continual structure infomax module could better preserve the previously learned knowledge. It clearly validates the effectiveness of our proposed continual structure infomax module in mitigating the catastrophic forgetting problem. As a whole, the superior performance clearly verifies the superiority of our proposed FCIL in learning novel class knowledge as well as mitigating the catastrophic forgetting problem. The accuracy changing curves with growing continual sessions in Fig. 2(a) further visually show the superiority of our FCIL. We further plot the accuracy changing curves on the CIFAR100 [22] and CUB200 [33] datasets separately in Fig. 2(b) and Fig. 2(c). It could be observed that our proposed FCIL also achieves the best performances on the two datasets, and it further ver-

Table 2. Overall ablation study of the proposed method on the miniImageNet dataset. Feature embedding infomax (FEI) and continual structure infomax (CSI) are the two functional modules we proposed.

Base	FEI	CSI	Accuracy in each session (%) ↑									KR	Avg
			1	2	3	4	5	6	7	8	9		
✓			75.87	70.22	65.92	62.64	59.96	57.11	54.58	52.66	51.49	67.86	61.16
✓	✓		76.34	71.18	67.16	63.85	61.05	58.23	55.33	53.67	52.32	68.54	62.13
✓		✓	75.87	70.16	65.76	63.76	60.34	57.52	54.96	53.06	52.06	68.61	61.39
✓	✓	✓	76.34	71.40	67.10	64.08	61.30	58.51	55.72	54.08	52.76	69.11	62.37



(a) FCIL (b) Baseline

Figure 3. Confusion matrix of the final classification results on the miniImageNet dataset [29]. (a) Baseline. (b) Our FCIL. We use red lines to partition the region of the base class and the new class. Our method improves the predictive power of the network which reduces the scattered confusion matrix.

ifies the superiority of our proposed feature embedding infomax and continual structure infomax modules in learning novel class knowledge and meantime mitigating the catastrophic forgetting problem.

4.3. Ablation study

All ablation studies are performed on the miniImageNet dataset [29] to evaluate the effectiveness of our proposed modules as well as some experimental settings.

Effectiveness of the proposed modules: We first conduct experiments to analyze the effectiveness of our proposed feature embedding infomax (FEI) and continual structure infomax (CSI) modules. The detailed results are listed in Table 2. In continual learning sessions ($t > 1$), the feature embedding infomax module is added to maximize the mutual information between the convolutional features produced by the fixed base backbone and the embedding features from the newly updated classifier. The performance improves over the baseline by 0.97% and 0.68% on Avg and KR, demonstrating that the proposed feature embedding infomax module could well transfer the strong learning capability of the base network into the following continual sessions to enhance the learning capability of the model on novel classes. It verifies the effectiveness of our proposed feature embedding the infomax module. The continual structure infomax learning module is conducted to maximize the mutual information between the current information structure and the previous ones. The performance promotion of the CSI model seems not significant in the early stages, since CSI is designed to alleviate catastrophic forgetting, and in the early stages (*i.e.,* t=2,3,4) the few new classes only pose limited influence on the learned knowledge. Further, with new classes increased in the later stages, the performance gain of CSI greatly increases. It achieves performance promotions of 0.75% and 0.23% separately on KR and Avg, which are clearly non-trivial. It shows that our proposed continual structure infomax module could well preserve the learned knowledge of the previous classes to mitigate the catastrophic forgetting issue. We believe it is because the attributes and class relationships could be well preserved during the continual sessions, and it verifies the effectiveness of the proposed continual structure infomax module. Finally, when both modules are applied, the performances could be further boosted to form a new state-of-the-art performance, quantitatively verifying the effectiveness of our proposed FCIL framework.

We further plot the confusion matrix of the baseline and our FCIL in Fig. 3. It could be observed that in the new classes, the diagonal of our FCIL has a darker color than the baseline model, vividly showing that the proposed FCIL is more powerful at learning new classes. In addition, it also indicates that the prediction distribution of FCIL on new classes is more concentrated than the baseline model, and it visually demonstrates the superiority of our FCIL in preserving learned knowledge.

Different methods for MI estimation and relation construction: The mutual information (MI) estimation ways can be adopted with the support of JSD [27] or KL [9], i.e., "MI_{JSD}" and "MI_{KL}". We achieve the information related in the continual structure infomax module with KNN or Cosine distance, named: "R_{KNN}" and "R_{Cos}". The FCIL results with the above different methods are listed in Table 3, and it could be observed that the performance of the four variants changes little. It means that it is the proposed structure preserve mechanism rather than the specifically chosen functions that mitigate the catastrophic forgetting problem. Therefore, we adopt KL to estimate the mutual information and use the Cosine distance to build the relationship between different classes in our work.

The level of convolutional Feature L: When performing the feature embedding infomax learning, we can estimate mutual information by using more layers of convolutional features $F^l_{conv}(x_i)$. To study the effect of different level sets, we construct three variants of FCIL by separately setting L as $\{1,2\}$,$\{1,3\}$, and $\{1,2,3\}$. The correspond-

Table 3. The results with different methods for MI estimation and relation construction on the miniImageNet dataset.

	Accuracy in each session (%) ↑									KR	Avg
	1	2	3	4	5	6	7	8	9		
$MI_{KL} + R_{KNN}$	76.34	71.40	67.07	64.05	61.29	58.51	55.73	54.07	52.74	68.08	62.35
$MI_{KL} + R_{Cos}$	76.34	71.40	67.10	64.08	61.30	58.51	55.72	54.08	52.76	69.11	62.37
$MI_{JSD} + R_{KNN}$	76.34	71.34	67.03	64.0	61.28	58.52	55.7	54.03	52.7	69.03	62.32
$MI_{JSD} + R_{Cos}$	76.34	71.26	66.99	64.03	61.24	58.46	55.65	53.98	52.63	68.94	62.28

Table 4. The comparison results between our FCIL and the open-source methods FACT [37] and CEC [35] by fixing the performance of the base classes on the miniImageNet dataset [29].

Methods	Accuracy in each session (%) ↑									KR↑	ΔFinal↑	Avg↑
	1	2	3	4	5	6	7	8	9			
CEC [35]	72.57	66.48	62.74	59.46	56.67	53.68	51.12	49.23	47.58	65.56	+1.16	57.73
FACT [37]	72.73	67.73	63.77	60.29	57.25	54.07	51.17	49.31	47.91	65.87	+0.83	58.25
FCIL(Ours)	72.63	67.71	63.70	60.67	57.61	54.73	52.09	50.26	48.74	66.94		58.71

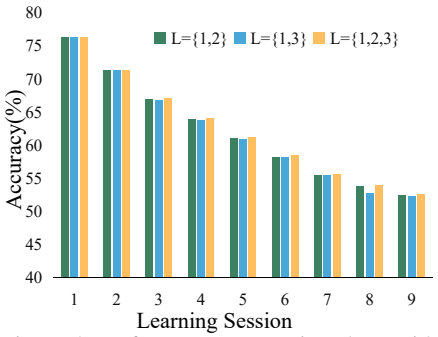


Figure 4. Performance comparison bars with different levels of convolutional features.

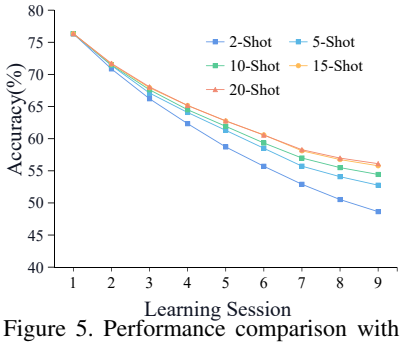


Figure 5. Performance comparison with the different shot numbers.

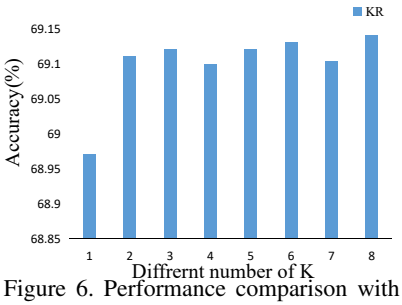


Figure 6. Performance comparison with different number of K.

ing accuracy changing bars are plotted in Fig. 4. It can be observed that the three variants change make little performance change. Further, it is worth noting that when employing all the levels of features (i.e., $L = \{1, 2, 3\}$), it could obtain better results than the other two variants, and all three convolutional features have been used. It means that our proposed feature embedding infomax module is able to continually transfer more useful encoding capabilities from multi-level convolutional features. It sufficiently validates the strong power of the feature embedding infomax module on transferring the encoding capability of the base network to the newly learned classifiers.

The shot number in the continual learning: We plot the accuracy changing curves with different shot numbers in Fig. 5. It shows that more sample shots could result in better few-shot continual classification performance. The reason is more shots could bring better class attribute distribution as well as more accurate class relations to benefit the learned new knowledge. Further, when the shot number is more than 15, the performance promotion becomes trivial. It shows that we only need a proper number of newly labeled samples to achieve satisfactory performances, verifying the effectiveness of our FCIL framework in addressing the few-shot continual classification task.

Different number of K: In the continual structure infomax learning module, we can use the different number of these previous information structures K to constrain the structure of learned knowledge in the continual stage. As K increases, there is a slight but smaller growth in the KR results, indicating that the structural information of the adjacent stages is adequately constrained for the current stage. This result can clearly show that it is possible to better mitigate catastrophic forgetting by imposing structural constraints on learned knowledge, but the demand for computational complexity and memory would increase significantly. Therefore, considering the performance and complexity of our method, we set K equal to 2 in our work.

Different methods with similar starting performance: As shown in Table 4, we give the comparison results of the FCIL, FACT [37], and CEC [35] methods with similar performance in the initial/base stage. We can observe that the three metrics of the FCIL framework are superior to the other methods, which indicates that the FCIL framework can retain the learned knowledge better, and the improvement is not because of the base model performance but our introduced FCIL. It clearly verifies the superiority of the proposed FCIL framework in effectively learning new knowledge and mitigating catastrophic forgetting.

5. Conclusion

In this paper, we propose a Few-show Continual Informal Learning (FCIL) framework to address the few-shot class continual learning task. Our FCIL enables a deep model to continually/incrementally learn new concepts from a few labeled samples and meantime prevent the model from forgetting the previously learned knowl-

edge. Inspired by the transfer entropy concept, a feature embedding infomax module is proposed by maximizing the mutual information between different level feature distributions and then grants the model to transfer the strong encoding capability of the base network into the new classes. Further, a continual structure infomax module is proposed to maximize the mutual information across these continually changing relations of the classes to resemble the continual learning progress, so that catastrophic forgetting could be well mitigated.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grants Nos. 61972204, 62072244), the fundamental research funds for the central universities under Grant 30919011232, the Natural Science Foundation of Shandong Province (Grant Nos. ZR2020LZH008, ZR2022LZH003), and in part by State Key Laboratory of High-end Server & Storage Technology.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 2

[2] 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. 2, 3

[3] Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T Lizier. Transfer entropy. In *An introduction to transfer entropy*, pages 65–95. 2016. 3, 4

[4] 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. 6

[5] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[6] Ali Cheraghian, Shafin Rahman, Sameera Ramasinghe, Pengfei Fang, Christian Simon, Lars Petersson, and Mehrtash Harandi. Synthesized feature based few-shot class-incremental learning on a mixture of subspaces. In *IEEE International Conference on Computer Vision*, pages 8661–8670, 2021. 1

[7] 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. 1, 2, 6

[8] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. 2021. 1

[9] 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. 2, 4, 7

[10] Howard Eichenbaum. How does the brain organize memories? *Science*, 277(5324):330–332, 1997. 2, 3, 4

[11] Yiduo Guo, Bing Liu, and Dongyan Zhao. Online continual learning through mutual information maximization. In *International Conference on Machine Learning*, pages 8109–8126. PMLR, 2022. 1, 3

[12] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and pattern recognition*, pages 770–778, 2016. 1

[14] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. 2022. 2

[15] 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. 6

[16] 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. 2, 3

[17] 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. 6

[18] Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[19] Baoyu Jing, Chanyoung Park, and Hanghang Tong. Hdmi: High-order deep multiplex infomax. In *Proceedings of the Web Conference 2021*, pages 2414–2424, 2021. 2

[20] Jayateja Kalla and Soma Biswas. S3c: Self-supervised stochastic classifiers for few-shot class-incremental learning. In *Proceedings of the European Conference on Computer Vision*, pages 432–448. Springer, 2022. 1, 2

[21] Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *National Academy of Sciences*, 111(9):3354–3359, 2014. 2

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 5, 6

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of The ACM*, 2012. 1

[24] 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. 6

[25] Yiqiao Mao, Xiaoqiang Yan, Qiang Guo, and Yangdong Ye. Deep mutual information maximin for cross-modal clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8893–8901, 2021. 2

[26] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2337–2345, 2021. 1, 2

[27] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016. 2, 7

[28] 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. 1, 6

[29] 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. 5, 6, 7, 8

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014. 1

[31] 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. 1, 2, 5, 6

[32] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 6

[33] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5, 6

[34] 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. 6

[35] 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. 1, 2, 6, 8

[36] Wenting Zhao, Gongping Xu, Zhen Cui, Siqiang Luo, Cheng Long, and Tong Zhang. Deep graph structural infomax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4920–4928, 2023. 3

[37] 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. 1, 2, 6, 8

[38] 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. 1, 2