Socialized Learning: Making Each Other Better Through Multi-Agent Collaboration

Xinjie Yao [*1 2 3] **Yu Wang** [*1 2 3] **Pengfei Zhu** [1 2 3] **Wanyu Lin** [4] **Jialu Li** [1 2 3] **Weihao Li** [5] **Qinghua Hu** [1 2 3]

Abstract

Learning new knowledge frequently occurs in our dynamically changing world, *e.g.*, humans culturally evolve by continuously acquiring new abilities to sustain their survival, leveraging collective intelligence rather than a large number of individual attempts. The effective learning paradigm during cultural evolution is termed socialized learning (SL). Consequently, a straightforward question arises: Can multi-agent systems acquire more new abilities like humans? In contrast to most existing methods that address continual learning and multi-agent collaboration, our emphasis lies in a more challenging problem: We prioritize the knowledge in the original expert classes, and as we adeptly learn new ones, the accuracy in the original expert classes stays superior among all in a directional manner. Inspired by population genetics and cognitive science, leading to unique and complete development, we propose Multi-Agent Socialized Collaboration (MASC), which achieves SL through interactions among multiple agents. Specifically, we introduce collective collaboration and reciprocal altruism modules, organizing collaborative behaviors, promoting information sharing, and facilitating learning and knowledge interaction among individuals. We demonstrate the effectiveness of multi-agent collaboration in an extensive empirical study. Our code will be publicly available at https://github.com/yxjdarren/SL.

[*]Equal contribution [1]College of Intelligence and Computing, Tianjin University, Tianjin, China [2]Engineering Research Center of City Intelligence and Digital Governance, Ministry of Education of the People's Republic of China, Tianjin, China [3]Haihe Lab of ITAI, Tianjin, China [4]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China [5]Department of Computer Science, Boston University, Boston, United States. Correspondence to: Pengfei Zhu <zhupengfei@tju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Every intelligent agent aspires to grow through learning new knowledge. Although existing continual learning (CL) paradigms are capable of acquiring new knowledge, they heavily rely on a substantial amount of data. Looking back on the history of human development, cultural evolution is accomplished by continuously learning new abilities beyond existing vital abilities for maintaining one's own survival (Henrich, 2016; Laland, 2017). In this process, new abilities are usually acquired by seeking guidance from experienced experts with their knowledge rather than through a large number of attempts and the collection of relevant examples. Such an effective learning paradigm is termed socialized learning (SL), which has been studied in cognitive science (Thompson et al., 2022).

Among the paradigms in machine learning, CL is the most relevant one for SL. It addresses the challenges of obtaining new knowledge while retaining old one over time (De Lange et al., 2021; Masana et al., 2022). However, CL focuses on an individual agent and seriously relies on large-scale training samples, making it difficult to learn as effectively as humans. Another alternative is federated learning (FL), which is a distributed learning paradigm (Zhang et al., 2021b; Li et al., 2023a). Compared to CL, FL acquires knowledge from multiple clients through aggregating different agents, leveraging knowledge in multiple agents. FL is primarily designed to integrate the knowledge of agents working on similar tasks, making it highly sensitive to data heterogeneity. This sensitivity is particularly pronounced in scenarios where there are significant differences in categories and knowledge among multiple agents (Shi et al., 2023). Therefore, existing learning paradigms are limited in realizing effective collective learning.

The existing paradigms struggle to learn new knowledge through interaction like humans. Due to diverse environments, individuals inherently exhibit significant heterogeneity. For models, the heterogeneity in data and models can lead to dimensional collapse when interactions are limited to the parameter space. Complete avoidance of data interaction is not necessary in many real-world scenarios without strong data privacy settings, such as emergency rescue situations. However, data interaction requires high com-

munication costs, making methods like FL, which involves repeated uploading and downloading, impractical. By revisiting existing paradigms, we find that two problems are still open: (1) Learning without relying on a large amount of data in a scenario with strong heterogeneity. (2) Learning with model interaction at low cost.

To address these challenges, we analyze two paradigms: CL and FL. As shown in Figure 1, relying solely on data for learning new knowledge, when the knowledge space is already populated by expert classes, can result in interference and decreased performance of those expert classes. Besides, when multiple agents learn through parameter transmission, the heterogeneity in data and models can lead to dimension collapse and overall performance decline. A solution to enable agents to learn from others at a low cost is SL, involving collaboration and knowledge interaction among multiple agents. Thus, two significant issues should be explored to realize it:

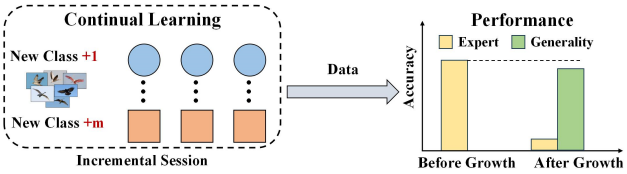
1: *How to establish sociability for collaboration?*

2: *How to leverage collective intelligence for learning?*

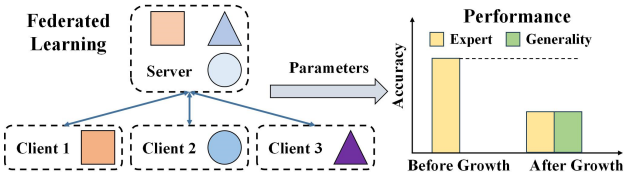
Parallel to perspectives in cognitive science (Mesoudi, 2021), this entails learning new knowledge from diverse experts and integrating it with individual needs to achieve growth. Motivated by this target, we employ multi-agent collaboration to imbue the model with versatility and directionality, *i.e.*, effectively learning a broader range of new general classes while directionally retaining the performance of original expert classes. When reflecting on the connection between sociality and cognitive function in humans, most cultural evolution has been driven by population genetics and cognitive science (Mesoudi, 2021). Inspired by this, we are strongly motivated to combine population-genetic-style modeling with directionally biased transformations. Specifically, in the context of multi-agent SL, we focus on acquiring additional capabilities by learning from various experts through SL. This involves shaping priors based on information received from other agents and ultimately enabling effective learning of new general classes while directionally maintaining the accuracy of the original expert classes within this SL paradigm.

To verify this premise, we turn to the design and analysis of a framework based on SL. First, we train a student agent by collective collaboration with multiple teacher agents, each of which is proficient in its own unique expert classes. Benefiting from both direct experiences obtained from samples and indirect experiences obtained from teachers, the student grows from a vanilla agent to a generalist, indicating its ability to classify all classes. Subsequently, each teacher, motivated by the desire to grow up, undergoes reciprocal altruism from the student, signifying that each teacher learns novel general classes. Finally, the grown teachers select the classifier based on Helmholtz free energy (Liu et al., 2020), enabling them to predict precisely. This process ensures the directional maintenance of their original expert classes while effectively learning new general classes. The contributions can be further detailed as follows:

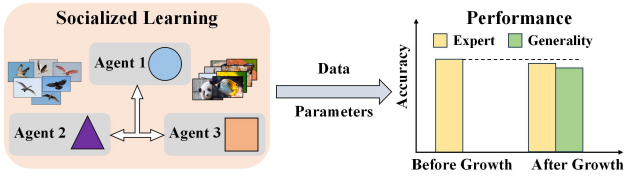
- We introduce a practical learning paradigm, socialized learning (SL), where multiple agents achieve their individual growth through collaborative interactions.
- We discuss SL using an information-based theoretical framework, exploring the impact of sociability information on SL capabilities.
- We propose a novel insight into the methodology of SL. Knowledge interaction occurs through collective collaboration, followed by reciprocal altruism, ensuring directional trade-offs among diverse abilities.



(a) Continual learning paradigm.



(b) Federated learning paradigm.



(c) Socialized learning paradigm.

Figure 1: Comparison of different learning paradigms.

2. Related Work

Continual learning (CL), a paradigm addressing the trade-off between stability and plasticity, requires an individual agent to retain previously acquired abilities while learning new ones effectively. Existing approaches (De Lange et al., 2021) can be broadly categorized into three main groups: Rehearsal-based techniques involve the selective replay of a subset of historical exemplars from old classes, either

sourced from prior tasks or generated using generative models (Rebuffi et al., 2017; Zhou et al., 2023; Zhao et al., 2020; Douillard et al., 2020; Lin et al., 2023; Tiwari et al., 2022; Yan et al., 2022). Regularization-based methods enhance the preservation of prior knowledge by introducing supplementary regularization terms into the loss function (Li & Hoiem, 2017; Pelosin et al., 2022; Zhou et al., 2022b; Liu et al., 2023; Kirkpatrick et al., 2017). Parameter isolation techniques allocate distinct parameter sets for each task, thereby averting interference between new and old knowledge during incremental learning (Yan et al., 2021; Wang et al., 2022a; Aljundi et al., 2017; Liu et al., 2021; Yang et al., 2022; Zhang et al., 2021a; Zhou et al., 2022a).

Existing CL works are mainly based on an individual agent without considering collaborative interactions among multiple agents. Although CL strives to retain old expert classes while learning new general ones, the impact of catastrophic forgetting inevitably leads to a decline in the performance of the initially mastered expert classes. In CL, the key challenge lies in the trade-off between stability and plasticity, *i.e.*, expert and generality, as illustrated in Figure 1a. SL leverages collaborative interactions among multiple agents to provide insight into the above issues.

Federated learning (FL), a paradigm of collaborative learning among multiple agents, is reliant on a centralized server overseeing the coordination of model training across a distributed network of devices. Depending on how the data is distributed in the feature and sample space, FL is divided into horizontal FL, vertical FL, and federated transfer learning (Li et al., 2023a). In horizontal FL, the datasets of different parties have the same feature space but little intersection on the sample space (McMahan et al., 2017; Wang et al., 2020; Li et al., 2021; Reddi et al., 2021; Li & Zhan, 2021; Zhang et al., 2022; Qu et al., 2022; Shi et al., 2023; Jhunjhunwala et al., 2023). Multiple parties with different features about the same set of users jointly train machine learning models in vertical FL (Liu et al., 2024; 2022; Castiglia et al., 2023). Federated transfer learning is an effective solution when data partitioning among parties involves a hybrid of horizontal and vertical partitioning (Liu et al., 2018; Wu & Zhang, 2023; Guo et al., 2023; Qi et al., 2023). Moreover, several works are integrations of FL and CL, wherein individual clients engage in a sequential acquisition of knowledge from distinct private data streams (Yoon et al., 2021; Dong et al., 2022).

Notably, a limitation of FL lies in its sensitivity to data heterogeneity. If attempting to induce the server to learn new classes by adding a client containing these classes, the strong heterogeneity of this client compared to others can result in a server dimension collapse in the model. Additionally, FL emphasizes privacy protection, which restricts its modes of collaborative interaction, impeding the learning of new general classes, as shown in Figure 1b. SL utilizes the directional transfer of data and knowledge among multiple agents to alleviate the above problems.

3. Sociability in Collaboration

In this section, to address the first issue highlighted in Section 1, we provide detailed definitions for the problem setup, followed by a concise theoretical analysis.

Problem setup: Let $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_N\}$ denote the N-agent set. X is a input space and Y is a label space. The unique data of the n-th agent is denoted as $\mathcal{D}_{\mathcal{A}_n} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{M_{\mathcal{A}_n}}$, where $M_{\mathcal{A}_n}$ represents the total number of samples for $\mathcal{A}_n$. $\boldsymbol{x}_i \in \mathbb{R}^D$ is a sample of class $y_i \in Y_{\mathcal{A}_n}$. To illustrate clearly, we take two agents, $\mathcal{A}_1$ and $\mathcal{A}_2$, as examples. For any two agents, $\mathcal{A}_1$ and $\mathcal{A}_2$, the sets of expert classes they possess are unique, *i.e.*, $\mathcal{D}_{\mathcal{A}_1}$ and $\mathcal{D}_{\mathcal{A}_2}$ contain expert classes that have not been seen from each other: $Y_{\mathcal{A}_1} \cap Y_{\mathcal{A}_2} = \varnothing$. After SL, both agents $\mathcal{A}_1$ and $\mathcal{A}_2$ can infer the classes $Y_{\mathcal{A}_1} \cup Y_{\mathcal{A}_2}$. Additionally, each agent maintains superior performance in its respective expert classes, *i.e.*, $\underset{k\in\{\mathcal{A}_1,\mathcal{A}_2\}}{\operatorname{argmin}} (\mathbb{E}_{x_k\sim P_{X_k}}[1 - \max_{y\in Y_{\mathcal{A}_1}} P(Y_{\mathcal{A}_1} = y|x_k)]) = \mathcal{A}_1$, $\underset{k\in\{\mathcal{A}_1,\mathcal{A}_2\}}{\operatorname{argmin}} (\mathbb{E}_{x_k\sim P_{X_k}}[1 - \max_{y\in Y_{\mathcal{A}_2}} P(Y_{\mathcal{A}_2} = y|x_k)]) = \mathcal{A}_2$. For $\mathcal{A}_1$ and $\mathcal{A}_2$, the newly learned general classes, corresponding to $Y_{\mathcal{A}_2}$ and $Y_{\mathcal{A}_1}$, respectively, should be learned as well as possible while maintaining the performance of the original expert classes directionally, *i.e.*, aspiring to possess both versatility and directionality, as illustrated in Figure 1c.

Definition 3.1. (Versatility) Each agent effectively learns a broader range of new general classes, *i.e.*, $\mathcal{A}_1$ and $\mathcal{A}_2$ can infer the classes $Y_{\mathcal{A}_1} \cup Y_{\mathcal{A}_2}$ with precision.

Definition 3.2. (Directionality) Each agent effectively learns new general classes while directionally maintaining its superior performance in its respective expert classes, *i.e.*, $\underset{k\in\{\mathcal{A}_1,\mathcal{A}_2\}}{\operatorname{argmin}} (\mathbb{E}_{x_k\sim P_{X_k}}[1 - \max_{y\in Y_{\mathcal{A}_1}} P(Y_{\mathcal{A}_1} = y|x_k)]) = \mathcal{A}_1$, $\underset{k\in\{\mathcal{A}_1,\mathcal{A}_2\}}{\operatorname{argmin}} (\mathbb{E}_{x_k\sim P_{X_k}}[1 - \max_{y\in Y_{\mathcal{A}_2}} P(Y_{\mathcal{A}_2} = y|x_k)]) = \mathcal{A}_2$.

To clearly elucidate the significance of multi-agent collaboration, we provide a definition of sociability information and subsequently conduct an analysis based on this premise:

Definition 3.3. (Sociability Information) For the input variables $X_{\mathcal{A}_1}$, $X_{\mathcal{A}_2}$, and the target Y, the sociability information provided by the agents can be defined as:

$$\Phi_{X_{\mathcal{A}_1}} = I(X_{\mathcal{A}_1}; Y|X_{\mathcal{A}_2}), \tag{1}$$

$$\Phi_{X_{\mathcal{A}_2}} = I(X_{\mathcal{A}_2}; Y|X_{\mathcal{A}_1}). \tag{2}$$

The $\Phi_{X_{\mathcal{A}_1}}$ and $\Phi_{X_{\mathcal{A}_2}}$ metrics quantify the sociability information possessed by $X_{\mathcal{A}_1}$ and $X_{\mathcal{A}_2}$ in SL, respectively.

Larger values of these metrics imply higher sociability information, indicating increased multi-agent collaboration. These measures collectively determine the collaborative efficiency and information sharing between $X_{\mathcal{A}_1}$ and $X_{\mathcal{A}_2}$. From the standard derivation in information theory, we can obtain the following relation:

$$I(X_{\mathcal{A}_1}, X_{\mathcal{A}_2}; Y) = \Phi_{X_{\mathcal{A}_1}} + \Phi_{X_{\mathcal{A}_2}} + I(X_{\mathcal{A}_1}; X_{\mathcal{A}_2}; Y). \quad (3)$$

Bayes error rate: The Bayes error rate (Fukunaga & Hummels, 1987) is introduced to measure agent performance, representing the lowest achievable error for any classifier or predictor from the multiple agents in inferring the target. Formally, considering two agent information $X_{\mathcal{A}_1}$ and $X_{\mathcal{A}_2}$, the Bayes errors for multi-agent and single-agent scenarios (assuming only $X_{\mathcal{A}_1}$ exists) in classification, denoted as $P_{e_c}^{mul}$ and $P_{e_c}^{sin}$ respectively, are defined as follows:

$$P_{e_c}^{mul} = \mathbb{E}_{x_{\mathcal{A}_1}, x_{\mathcal{A}_2} \sim P_{X_{\mathcal{A}_1}, X_{\mathcal{A}_2}}} [1 - \max_{y \in Y} P(Y = y | x_{\mathcal{A}_1}, x_{\mathcal{A}_2})], \quad (4)$$

$$P_{e_c}^{sin} = \mathbb{E}_{x_{\mathcal{A}_1} \sim P_{X_{\mathcal{A}_1}}} [1 - \max_{y \in Y} P(Y = y | x_{\mathcal{A}_1})]. \quad (5)$$

Theorem 3.4. *We build upon prior findings (Cover, 1999; Feder & Merhav, 1994; Li et al., 2023b) and position the Bayes error rates $P_{e_c}^{mul}$ and $P_{e_c}^{sin}$ centrally. For the variables $X_{\mathcal{A}_1}$, $X_{\mathcal{A}_2}$, and Y, the relationships are given by:*

$$\frac{H(Y|X_{\mathcal{A}_1}, X_{\mathcal{A}_2}) - \log 2}{\log |Y|} \leq P_{e_c}^{mul} \leq 1 - \exp(-H(Y \mid X_{\mathcal{A}_1}, X_{\mathcal{A}_2})), \quad (6)$$

$$\frac{H(Y|X_{\mathcal{A}_1}) - \log 2}{\log |Y|} \leq P_{e_c}^{sin} \leq 1 - \exp(-H(Y \mid X_{\mathcal{A}_1})). \quad (7)$$

Since

$$\Phi_{X_{\mathcal{A}_2}} = H(Y|X_{\mathcal{A}_1}) - H(Y|X_{\mathcal{A}_1}, X_{\mathcal{A}_2}). \quad (8)$$

We can derive

$$\frac{H(Y|X_{\mathcal{A}_1}) - \Phi_{X_{\mathcal{A}_2}} - \log 2}{\log |Y|} \leq P_{e_c}^{mul} \leq 1 - \exp(-H(Y \mid X_{\mathcal{A}_1}) + \Phi_{X_{\mathcal{A}_2}}). \quad (9)$$

Remark 3.5. The disparity between $P_{e_c}^{mul}$ and $P_{e_c}^{sin}$ reflects the degree of collaboration among multiple agents. In scenarios lacking sociability information ($\Phi_{X_{\mathcal{A}_2}} = 0$), the agent performs comparably under both settings. However, as $\Phi_{X_{\mathcal{A}_2}}$ increases, multi-agent performance increases. Due to the directionality of $\mathcal{A}_1$ and $\mathcal{A}_2$ ($\Phi_{X_{\mathcal{A}_2}} > 0$), the multi-agent performance is probably better than the single-agent performance. Collaboration among multiple intelligent agents can facilitate the growth of each agent.

4. Methodology

In this section, to address the second issue highlighted in Section 1, we seek to enhance versatility and directionality in SL. The training target is to prepare for learning new general classes while directionally maintaining expert classes. We achieve this through two aspects. On the one hand, to enable multiple agents to collaborate effectively, we aim to let multiple teacher agents (TAs) collectively guide a student agent (SA) as a medium for knowledge interaction. On the other hand, to facilitate the growth of each TA, meaning to maintain their expert accuracy while maximizing general accuracy, we seek the capability for TAs to be reciprocally educated by the SA. TAs that have been reciprocally educated are considered a form of growth, and these grown TAs can better guide the student, establishing a virtuous cycle.

We first introduce the SL framework and then discuss how to enable multiple agents to collaborate effectively and how to facilitate the growth of agents.

4.1. Socialized Learning Framework

We take a closer look at the above two aspects, *i.e.*, collaboration and growth, aiming to find the key factor that connects them. Inspired by (Li et al., 2023b), we provide an information-theoretical analysis in the previous section for SL and explore the impact of sociability information on SL capabilities.

Based on the above analysis, we seek to design a unified SL framework $U(\cdot)$ with moderate multi-agent information interactions to utilize complementarity and obtain versatility.

$$U_i(\mathcal{A}_1, \mathcal{A}_2) = \psi(\varphi(\mathcal{A}_1, \mathcal{A}_2), \mathcal{A}_i), \quad (10)$$

where $\mathcal{A}_1$ and $\mathcal{A}_2$ denote two different agents, $i \in \{1, 2\}$ denotes the i-th agent in growth, $\varphi(\cdot)$ denotes the approach of collaboration, and $\psi(\cdot)$ denotes the approach of growth.

4.2. Multi-agent Socialized Learning

Driven by the above analysis and (Wang et al., 2022b; 2023a; Yang et al., 2021; Wang et al., 2023b), we aim to address SL from two views: collaboration and growth. We introduce a novel approach based on SL, referred to as Multi-Agent Socialized Collaboration (MASC), as depicted in Figure 2. MASC accomplishes SL through two primary modules: collective collaboration and reciprocal altruism. For clarity, we elaborate on each module contained within the agent.

Effect of collective collaboration: In MASC, the student obtains collective intelligence through collaborative distillation. The objective function is composed of the cross-entropy loss $\mathcal{L}_{ce}^s$, knowledge distillation loss $\mathcal{L}_{kd}^s$, and energy alignment loss $\mathcal{L}_{al}^s$, as follows:

$$\mathcal{L}^s = \mathcal{L}_{ce}^s(\boldsymbol{p}^s, \boldsymbol{p}^{gt}) + \lambda_1 \sum_{i=1}^{N} \mathcal{L}_{kd}^s(\hat{\boldsymbol{p}}^{s_i}, \boldsymbol{p}^{t_i}) + \lambda_2 \mathcal{L}_{al}^s(\boldsymbol{p}^s, \Delta), \quad (11)$$

where $\boldsymbol{p}^s$ denotes the prediction of the student, $\boldsymbol{p}^{gt}$ denotes the ground truth, $\hat{\boldsymbol{p}}^{s_i}$ denotes the prediction of the classification head of the teacher after receiving features from the student, $\boldsymbol{p}^{t_i}$ denotes the prediction of the teacher, and Δ is a hyper-parameter. λ_1 and λ_2 are the trade-off parameters.

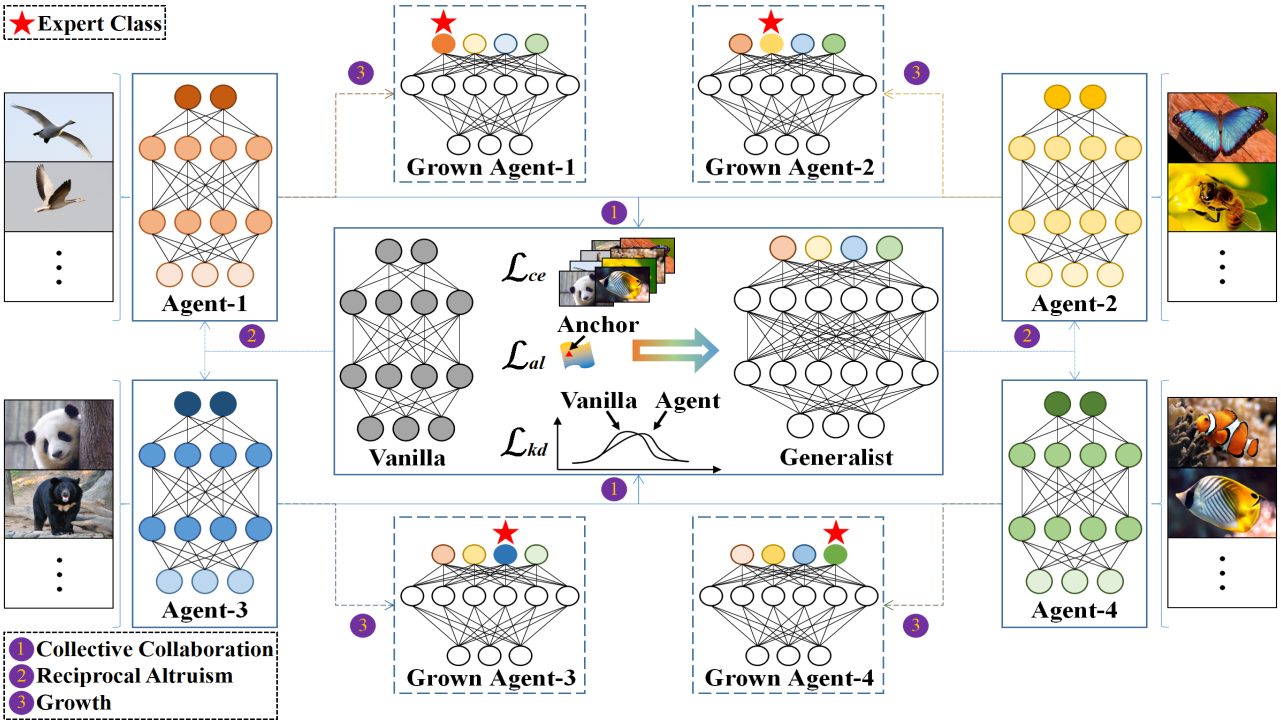


Figure 2: Illustration of MASC. Step ①: Collective Collaboration. A vanilla agent grows into a generalist by cross-entropy loss $\mathcal{L}_{ce}$, knowledge distillation loss $\mathcal{L}_{kd}$, and alignment loss $\mathcal{L}_{al}$. Step ②: Reciprocal Altruism. Through interactions with the generalist, each independent agent learns new general classes while directionally preserving their original expert classes. Step ③: Growth is achieved through the combined effects of collective collaboration and reciprocal altruism. The red stars ⋆ denote expert classes.

Since MASC involves N agents, there is a risk of confronting the conflict between ground-truth targets and distillation targets, *i.e.*, the distillation targets predicted by the teacher have a significant discrepancy with the ground-truth targets assigned to the student. This carefully designed socialized collaborative distillation approach allows $\mathcal{L}_{kd}$ to avoid affecting the weight updates of certain student heads, thereby mitigating the conflict between $\mathcal{L}_{ce}$ and $\mathcal{L}_{kd}$. To facilitate better knowledge interaction, we align the teacher and student using $\mathcal{L}_{al}^{s}$ to restrict their energies around the anchor Δ. Next, we will provide a detailed explanation of $\mathcal{L}_{al}^{s}$ in reciprocal altruism.

Effect of reciprocal altruism: Inspired by previous works (Liu et al., 2020; Wang et al., 2023b) related to Helmholtz free energy (HFE), we employ a divide and conquer approach to implement reciprocal altruism. Specifically, we obtain an anchor-based student through Eq. (11), which comprises a backbone $f_b^s(\cdot)$ and a classifier $f_c^s(\cdot)$. As the student grows up through collective collaboration, transitioning from a vanilla agent to a generalist, it acquires classification ability across all classes. Compared to an individual teacher, the student possesses knowledge of a wider range of classes. However, its classification ability in the classes where the teacher excels remains far below that of the teacher. A straightforward idea arises: Can the student reciprocally benefit the teacher, enabling the teacher to maintain its expert classes while learning novel general classes? To achieve this goal, we keep $f_b^s(\cdot)$ fixed and train a grown teacher classifier $f_c^{t_i'}(\cdot)$ under the guidance of the original teacher, as follows:

$$\mathcal{L}^{t_i'} = \mathcal{L}_{ce}^{t_i'}\left(\boldsymbol{p}^{t_i'}, \boldsymbol{p}^{gt}\right) + \lambda_1 \mathcal{L}_{kd}^{t_i'}\left(\hat{\boldsymbol{p}}^{t_i'}, \boldsymbol{p}^{t_i}\right) + \lambda_2 \mathcal{L}_{al}^{t_i'}\left(\boldsymbol{p}^{t_i'}, \Delta\right), \quad (12)$$

where $\boldsymbol{p}^{t_i'}$ denotes the prediction of the grown teacher, $\hat{\boldsymbol{p}}^{t_i'}$ denotes the prediction of the classification head of the original teacher after receiving features from the grown teacher. $\boldsymbol{p}^{t_i}$, $\boldsymbol{p}^{gt}$, Δ, λ_1 and λ_2 are the same in Eq. (11).

The grown teacher consists of $f_b^s(\cdot)$, $f_c^s(\cdot)$, and $f_c^{t_i'}(\cdot)$. During inference, the agent will transfer the features extracted by $f_b^s(\cdot)$ to both $f_c^s(\cdot)$ and $f_c^{t_i'}(\cdot)$ to obtain their respective HFE, and we employ the classifier with a higher HFE for prediction. Next, we define the normal energy function for a given input-label pair $(\boldsymbol{x}, y)$ as follows:

$$E^k(\boldsymbol{x}, y) = -h^k(\boldsymbol{x})[y], \quad (13)$$

where $h^k(\boldsymbol{x}) = f_c^k(f_b^s(\boldsymbol{x}))$ is the logits of the k-th classifier, and $h^k(\boldsymbol{x})[y]$ is the logit value of $y \in Y_k$. Then, HFE can

be defined as follows:

$$\mathcal{F}^k(\boldsymbol{x}) = -\log \sum_{y \in Y_k} \exp\left(-E^k(\boldsymbol{x}, y)\right). \tag{14}$$

To align the HFE of different classifiers in the same space, we employ $\mathcal{L}_{al}^s$, which constrains the HFE of each classifier with a fixed anchor Δ, as follows:

$$\mathcal{L}_{al}^k = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_k}\left(\mathcal{F}^k(\boldsymbol{x}) - \Delta\right)^2. \tag{15}$$

Inference of SL: Under the influence of collective collaboration and reciprocal altruism, agents have acquired the conditions for SL. To achieve the simultaneous maintenance of performance in expert classes and high performance in general classes, we need to utilize the general classifier $f_c^s(\cdot)$ and expert classifier $f_c^{t_i'}(\cdot)$. The final prediction is made by obtaining the classifier with the highest HFE, as follows:

$$k^* = \underset{k \in \{s, t_i'\}}{\operatorname{argmax}}(-\mathcal{F}^k(\boldsymbol{x})). \tag{16}$$

Then, the final prediction can be obtained as:

$$y^* = f_c^{k^*}(f_b^s(\boldsymbol{x})). \tag{17}$$

For a clearer understanding of training, we have described the algorithm in Algorithm 1.

Algorithm 1 Training for MASC.

Input: Datasets $\mathcal{D}_s = (\mathcal{D}_{t_1}, \cdots, \mathcal{D}_{t_N})$, teacher numbers N, energy anchor Δ;
Output: Backbone $f_b^s(\cdot)$, student classifier $f_c^s(\cdot)$, teacher classifier $f_c^{t_i'}(\cdot)$;

```
Randomly initialize f_b^s(·), f_c^s(·), f_c^{t_i'}(·);
while not converged do
    Get a mini-batch of training data from D_s;
    Calculate the student loss L^s by Eq. (11);
    Update the student agent, i.e., f_b^s(·) and f_c^s(·);
end while
for i = 1, ···, N do
    while not converged do
        Get a mini-batch of training data from D_{t_i};
        Calculate the grown teacher loss L^{t_i'} by Eq. (12);
        Update the i-th grown teacher agent, i.e., f_c^{t_i'}(·);
    end while
end for
```

5. Experiment

In this section, we compare MASC on CIFAR10 and CIFAR100 datasets with state-of-the-art methods. We discuss the methods of data-driven knowledge distillation, and the ablations verify the effectiveness of collective collaboration and reciprocal altruism. More experiments and details of implementation refer to supplementary materials.

5.1. Implementation Details

Dataset split: We use two versions of dataset split. The first one involves dividing CIFAR10 evenly among 5 agents, with each agent having 2 classes (abbreviated as expert-class, general-class-1, general-class-2, general-class-3, and general-class-4), *i.e.*, CIFAR10-5-2, while CIFAR100 is split evenly among 4 agents, each having 25 classes, *i.e.*, CIFAR100-4-25. Similarly, the second version is CIFAR10-2-5 and CIFAR100-5-20. In the following experiments, we primarily employ the first version as the default scheme, and the experiments with the second version are presented in the supplementary materials.

Compared methods: We first compare to CL methods DER (Yan et al., 2021), EWC (Kirkpatrick et al., 2017), FOSTER (Wang et al., 2022a), iCaRL (Rebuffi et al., 2017), LwF (Li & Hoiem, 2017), MEMO (Zhou et al., 2023), PODNet (Douillard et al., 2020), and WA (Zhao et al., 2020). Besides, we also compare to FL methods FedAvg (McMahan et al., 2017), FedAvgM (Hsu et al., 2019), FedNova (Wang et al., 2020), FedDecorr (Shi et al., 2023), FedProx (Li et al., 2020), FedSAM (Qu et al., 2022), MOON (Li et al., 2021), and GLFC (Dong et al., 2022).

5.2. Versatility and Directionality are All You Need

MASC is compared with two paradigms: CL and FL. We report the performance on the CIFAR10 and CIFAR100 datasets, as shown in Tables 1 and 2.

Method	Accuracy before growth			Accuracy after growth		
	Expert	General	Average	Expert	General	Average
DER (Yan et al., 2021)	96.67	0.00	48.34	64.70	74.13	69.41
EWC (Kirkpatrick et al., 2017)	96.54	0.00	48.27	46.00	64.08	55.04
FOSTER (Wang et al., 2022a)	96.69	0.00	48.35	64.90	71.88	68.39
iCaRL (Rebuffi et al., 2017)	96.57	0.00	48.29	46.45	64.55	55.50
LwF (Li & Hoiem, 2017)	96.70	0.00	48.35	40.85	66.23	53.54
MEMO (Zhou et al., 2023)	96.60	0.00	48.30	78.50	61.46	69.98
PODNet (Douillard et al., 2020)	96.66	0.00	48.33	80.15	66.64	73.39
WA (Zhao et al., 2020)	96.72	0.00	48.36	63.90	74.00	68.95
FedAvg (McMahan et al., 2017)	96.02	0.00	48.01	91.50	63.13	77.31
FedAvgM (Hsu et al., 2019)	96.16	0.00	48.08	91.70	63.88	77.79
FedNova (Wang et al., 2020)	96.31	0.00	48.16	**91.90**	72.04	81.97
FedDecorr (Shi et al., 2023)	96.35	0.00	48.18	91.85	**75.54**	**83.69**
FedProx (Li et al., 2020)	96.27	0.00	48.14	91.40	70.95	81.18
FedSAM (Qu et al., 2022)	96.09	0.00	48.05	91.85	63.40	77.63
MOON (Li et al., 2021)	96.19	0.00	48.10	90.00	67.59	78.79
GLFC (Dong et al., 2022)	96.61	0.00	48.31	33.05	70.06	51.56
MASC	96.65	0.00	48.33	**93.40**$_{(+1.50)}$	**81.04**$_{(+5.50)}$	**87.22**$_{(+3.53)}$

Table 1: Comparison of detailed accuracy across different classes before and after growth on CIFAR10 dataset. The 1^{st}/2^{nd} best results are indicated in red/blue.

CL forgets original knowledge during growth: We observe that the general classes often exhibit outstanding performance in CL methods. However, the performance of

Method	Accuracy before growth			Accuracy after growth		
	Expert	General	Average	Expert	General	Average
DER (Yan et al., 2021)	66.45	0.00	33.23	54.04	54.53	54.29
EWC (Kirkpatrick et al., 2017)	66.12	0.00	33.06	19.25	43.93	31.59
FOSTER (Wang et al., 2022a)	66.41	0.00	33.21	34.40	53.32	43.86
iCaRL (Rebuffi et al., 2017)	66.25	0.00	33.13	31.48	46.45	38.97
LwF (Li & Hoiem, 2017)	66.38	0.00	33.19	15.48	52.00	33.74
MEMO (Zhou et al., 2023)	66.39	0.00	33.20	51.68	44.84	48.26
PODNet (Douillard et al., 2020)	66.34	0.00	33.17	54.36	41.92	48.14
WA (Zhao et al., 2020)	66.47	0.00	33.24	51.12	47.71	49.41
FedAvg (McMahan et al., 2017)	65.42	0.00	32.71	56.00	54.05	55.03
FedAvgM (Hsu et al., 2019)	65.48	0.00	32.74	58.56	54.16	56.36
FedNova (Wang et al., 2020)	66.00	0.00	33.00	58.32	54.64	56.48
FedDecorr (Shi et al., 2023)	66.12	0.00	33.06	**60.16**	**56.93**	**58.55**
FedProx (Li et al., 2020)	65.56	0.00	32.78	58.60	54.48	56.54
FedSAM (Qu et al., 2022)	66.08	0.00	33.04	59.68	55.41	57.55
MOON (Li et al., 2021)	65.29	0.00	32.65	50.12	47.11	48.61
GLFC (Dong et al., 2022)	66.33	0.00	33.17	47.04	40.49	43.77
MASC	66.28	0.00	33.14	**65.40**$_{(+5.24)}$	**58.64**$_{(+1.71)}$	**62.02**$_{(+3.47)}$

Table 2: Comparison of detailed accuracy across different classes before and after growth on CIFAR100 dataset. The 1^{st}/2^{nd} best results are indicated in red/blue.

expert classes tends to deteriorate rapidly. Naturally, a question arises: Can we consider the classes learned in the most recent session as expert classes? Firstly, such an assumption does not align with our setup. Secondly, even if we consider the most recently learned general classes as expert classes, the overall performance is still hampered by catastrophic forgetting. We aim for expert classes to be a stable set of classes throughout the learning process, rather than a degradation set of classes with learning progress.

FL loses sight of the focus among multiple agents: We observe that FL is the most competitive paradigm in terms of the accuracy of expert classes. However, we observe that the accuracy of expert and general classes in most FL methods is quite close. A straightforward question arises: Is this indicative of non-directional learning? Obviously, it indicates that FL does not prioritize the maintenance of expert-class performance. FL exhibits a degree of randomness and cannot learn based on pre-established preferences.

SL utilizes collective intelligence: MASC is a method designed based on the SL paradigm that combines the best of all worlds and divide-and-conquer. This method leverages the knowledge of collective intelligence to its fullest extent. Specifically, MASC not only learns knowledge through ground truth but also receives guidance from various teachers. This SL can enhance the ability of the agent. In addition, through reciprocal altruism and parameter isolation, the agent can learn more classes while maintaining expert accuracy. MASC not only enables one agent to grow up but enables every agent in the population to grow up, as shown in Table 3. We observe that this SL is directional and comprehensive.

Analysis of versatility and directionality: Versatility and directionality are crucial in SL. Possessing versatility implies that the agent can learn more general classes, while having directionality means maintaining the performance

Dataset	Method	Accuracy before growth			Accuracy after growth		
		Expert	General	Average	Expert	General	Average
CIFAR10	Agent-1	96.65	0.00	48.33	93.40	81.04	**87.22**
	Agent-2	86.20	0.00	43.10	86.15	68.08	**77.11**
	Agent-3	93.00	0.00	46.50	90.75	69.61	**80.18**
	Agent-4	96.85	0.00	48.43	92.10	81.19	**86.64**
	Agent-5	96.60	0.00	48.30	94.75	72.01	**83.38**
CIFAR100	Agent-1	66.28	0.00	33.14	65.40	58.64	**62.02**
	Agent-2	67.60	0.00	33.80	65.36	55.68	**60.52**
	Agent-3	64.84	0.00	32.42	61.64	58.59	**60.11**
	Agent-4	68.40	0.00	34.20	65.92	59.57	**62.75**

Table 3: Comparison of different agents before and after growth on CIFAR10 and CIFAR100 datasets.

of expert classes in a directional manner. We analyze versatility and directionality by comparing the average accuracy across all classes and the accuracy difference between expert classes and the best-performing general classes. As shown in Figure 3, most CL methods exhibit obvious forgetting in the performance of expert classes (negative values in orange bars), and most FL methods fail to maintain directionality in the performance of expert classes (orange bar values close to zero).

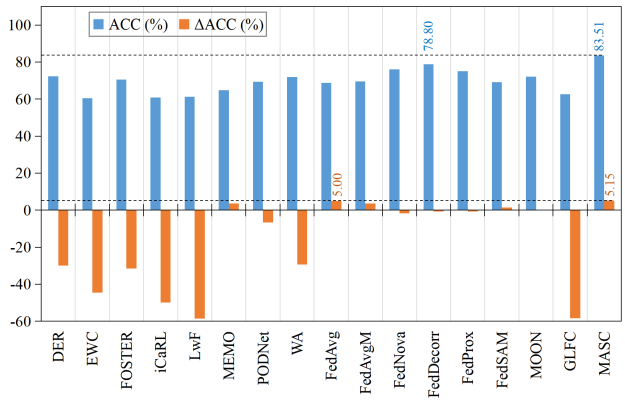
5.3. Data-driven Knowledge Distillation

Knowledge distillation encompasses various methods for extracting knowledge from agents. In practice, most existing machine learning methods are data-driven, so a straightforward idea arises: Can task-specific knowledge be distilled from data? Inspired by (Yang et al., 2021), we use unique data from each teacher to obtain the corresponding mean and variance (see supplementary material), thereby generating data to train agents without directly accessing the real samples. As shown in Figure 4, data-driven knowledge distillation contributes to agent performance.

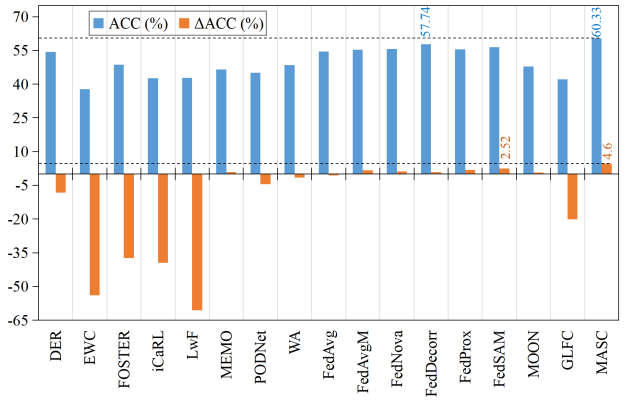
Dataset	Method	Accuracy before growth			Accuracy after growth		
		Expert	General	Average	Expert	General	Average
CIFAR10	Agent-1	89.40	0.00	44.70	84.00	65.31	**74.66**
	Agent-2	78.62	0.00	39.31	75.70	49.53	**62.61**
	Agent-3	77.60	0.00	38.80	72.60	61.08	**66.84**
	Agent-4	89.60	0.00	44.80	86.15	57.05	**71.60**
	Agent-5	87.70	0.00	43.85	83.60	64.26	**73.93**
CIFAR100	Agent-1	44.16	0.00	22.08	43.52	30.99	**37.25**
	Agent-2	42.08	0.00	21.04	41.28	31.24	**36.26**
	Agent-3	46.88	0.00	23.44	42.56	32.17	**37.37**
	Agent-4	45.04	0.00	22.52	44.08	31.80	**37.94**

Table 4: Comparison of different agents before and after growth on generated CIFAR10 and CIFAR100 datasets.

Since different agents have diverse classes, they can explore data-driven knowledge unique to their data, facilitating agent-specific knowledge interaction. To further validate the effectiveness of data-driven knowledge distillation, we conduct extensive experiments, as illustrated in Table 4. We have observed that data-driven knowledge distillation not

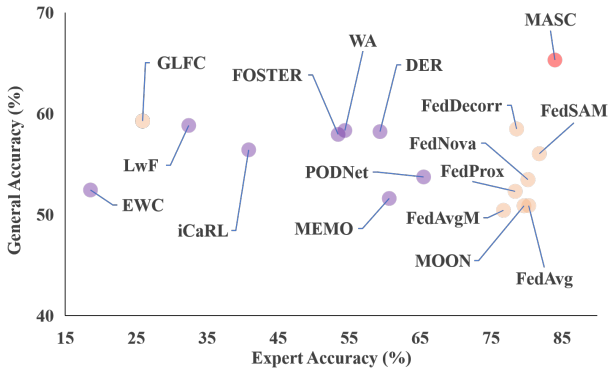


(a) Comparison on CIFAR10 dataset.

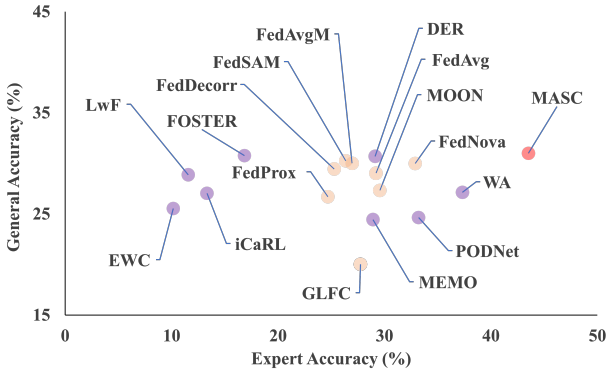


(b) Comparison on CIFAR100 dataset.

Figure 3: Analysis of versatility and directionality on CIFAR10 and CIFAR100. Blue bars denote the average accuracy across all classes, while orange bars denote the difference in accuracy between expert classes and the best-performing general classes. Blue and orange bars correlate positively with versatility and directionality, respectively.



(a) Comparison on generated CIFAR10 dataset.



(b) Comparison on generated CIFAR100 dataset.

Figure 4: Performance comparison after growth. Different colors denote different learning paradigms.

only provides a novel insight into traditional knowledge distillation but also offers a potential solution to reduce data transmission overhead.

5.4. Ablation Study

To further verify the significance of each module in MASC, we conduct the ablation study as shown in Table 5.

When using only collective collaboration (CC), we observe that the performance of expert classes and general classes is very close, with instances where general classes even outperform. This indicates that the CC module is a non-directional interaction primarily designed to facilitate the learning of new general classes in a multi-agent collaborative setting without ensuring the performance of the original expert classes. When using only reciprocal altruism (RA), we observe a significant improvement in the performance of expert classes over general classes, but poorer overall accuracy for both. This is attributed to the dominance of the expert classifier, as the majority of predictions are made using the expert classifier, failing to achieve a divide-and-conquer strategy. Finally, when combining CC and RA, we find that the agent achieves the expected outcome, maintaining the performance of the original expert classes while effectively learning new general classes.

Dataset	CC	RA	Expert classes	General classes	Average
CIFAR10			44.90	82.30	63.60
		✓	97.00	19.75	58.38
	✓	✓	93.40	81.04	**87.22**
CIFAR100			62.16	61.55	61.85
		✓	67.92	54.36	61.14
	✓	✓	65.40	58.64	**62.02**

Table 5: Ablation study on CIFAR10 and CIFAR100. In the table, "✓" denotes MASC with the module.

5.5. Discussion

We expound upon the concept of SL, a paradigm rooted in the principles of collaboration and knowledge interaction among multiple agents. SL establishes a foundational system where agents interact, co-learn, and grow collaboratively, enhancing individual abilities. Table 6 illustrates distinct characteristics of various learning paradigms when facing different challenges. SL, in particular, exhibits the following traits:

i) Cooperativity: Multiple agents engage in interactive collaboration, leveraging collective intelligence.

ii) No heterogeneity: Due to factors such as data distribution, heterogeneity is inevitable, and the ability to resist heterogeneity holds practical significance.

iii) Versatility: With the emergence of new classes, the agent needs to learn more general classes while preserving the performance of the original expert classes.

iv) Directionality: Due to diverse environments and personalized requirements, possessing the directionality for targeted growth provides a novel insight.

Paradigm	Cooperativity	No heterogeneity	Versatility	Directionality
FL	✓	×	×	×
CL	×	✓	✓	×
SL	✓	✓	✓	✓

Table 6: Comparison of different learning paradigms.

Based on the aforementioned traits, SL has the following four advantages:

i) Enhanced synergy: SL promotes a superior level of synergy among agents by leveraging the collective strengths and capabilities of the group. This design optimizes the overall learning outcome, surpassing the results achievable through individual agents teaching each other.

ii) Scalability: The inherent scalability of SL facilitates the inclusion of an increasing number of agents without the complexity associated with direct, pairwise teaching-learning interactions. This feature is pivotal for large-scale applications, ensuring the approach adapts seamlessly to expanding agent networks.

iii) Flexibility and robustness: The paradigm of SL, which does not rely on direct teaching among agents, enhances robustness to the variability in performance and knowledge of individual agents. This ensures the collaborative learning process remains resilient, effectively mitigating potential inaccuracies or inefficiencies from individual contributions.

iv) Efficient resource utilization: Designed for the efficient use of computational and communication resources, SL minimizes the necessity for extensive data exchange among agents. Instead, it focuses on the collective generation and dissemination of learning updates, optimizing resource allocation.

SL is not an independent learning paradigm, and it provides insights into several paradigms. For addressing the stability-plasticity dilemma, CL may leverage collective intelligence, with one subset of agents focusing on stability and another on plasticity. To tackle the challenge of heterogeneity, FL can explore multi-agent data-driven knowledge distillation and collective collaboration. Additionally, SL offers insights into other learning paradigms, such as long-tail distribution learning and transfer learning. For handling head-and-tail classes, a divide-and-conquer approach with collaborative interaction can be employed. To bridge the gap between the source domain and the target domain, multiple agents can aggregate shared knowledge, which is relatively universal. This enables the transfer of learned knowledge from related tasks, enhancing learning for new tasks.

6. Conclusion

In this paper, we introduce a practical SL paradigm with rigid mathematical motivation and explanation rooted in information theory. Additionally, we carefully devise MASC based on SL, aiming to effectively learn new general classes while directionally preserving the performance of the original expert classes, *i.e.*, learn from others and be itself. In essence, we treat expert classes and general classes as two sides of the same coin rather than two separate problems. Such a unified framework for collective collaboration and reciprocal altruism provides a novel insight into SL. In our future work, we plan to delve further into the potential of combining multiple modalities with multiple agents.

Impact Statement

In real-world scenarios, the collection of the SL dataset requires many different classes of samples, and multiple agents are also used to collect the samples, which means that the lives of some plants and animals will be disturbed during the data collection process. Therefore, it is necessary to avoid disturbing or affecting the normal life of plants and animals as much as possible when collecting the SL dataset to prevent disrupting the ecological balance. Possible limitations of this work include the absence of the number of classes in the real world for the datasets. The number of classes in the existing SL dataset is limited.

In the future, we will consider the dynamic relationships of different classes under different agents, exploiting the complementarity of each other more sufficiently, and obtaining more abilities. Moreover, we aim to break the bottleneck of the SL dataset and construct a large-scale benchmark with diverse classes and modalities for exhaustive evaluation.

Acknowledgements

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0116500, in part by the National Natural Science Foundation of China under Grants 61925602, U23B2049, 62106174, 62222608, and 62266035, and in part by Tianjin Natural Science Funds for Distinguished Young Scholar under Grant 23JCJQJC00270.

References

Aljundi, R., Chakravarty, P., and Tuytelaars, T. Expert gate: Lifelong learning with a network of experts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7120–7129, 2017.

Castiglia, T., Zhou, Y., Wang, S., Kadhe, S., Baracaldo, N., and Patterson, S. LESS-VFL: communication-efficient feature selection for vertical federated learning. In *International Conference on Machine Learning*, volume 202, pp. 3757–3781. PMLR, 2023.

Cover, T. M. *Elements of Information Theory*. John Wiley & Sons, 1999.

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.

Dong, J., Wang, L., Fang, Z., Sun, G., Xu, S., Wang, X., and Zhu, Q. Federated class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10164–10173, 2022.

Douillard, A., Cord, M., Ollion, C., Robert, T., and Valle, E. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*, pp. 86–102. Springer, 2020.

Feder, M. and Merhav, N. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

Fukunaga, K. and Hummels, D. M. Bayes error estimation using parzen and k-nn procedures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):634–643, 1987.

Guo, J., Ho, I. W. H., Hou, Y., and Li, Z. Fedpos: A federated transfer learning framework for csi-based wi-fi indoor positioning. *IEEE Systems Journal*, 17(3):4579–4590, 2023.

Henrich, J. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press, 2016.

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Jhunjhunwala, D., Wang, S., and Joshi, G. Fedexp: Speeding up federated averaging via extrapolation. In *International Conference on Learning Representations*, 2023.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

Laland, K. N. *Darwin's unfinished symphony: how culture made the human mind*. Princeton University Press, 2017.

Langley, P. Crafting papers on machine learning. In *International Conference on Machine Learning*, pp. 1207–1216. PMLR, 2000.

Li, Q., He, B., and Song, D. Model-contrastive federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10713–10722, 2021.

Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., and He, B. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3347–3366, 2023a.

Li, S., Du, C., Zhao, Y., Huang, Y., and Zhao, H. What makes for robust multi-modal models in the face of missing modalities? *arXiv preprint arXiv:2310.06383*, 2023b.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Li, X. and Zhan, D. Fedrs: Federated learning with restricted softmax for label distribution non-iid data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 995–1005, 2021.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Lin, H., Zhang, B., Feng, S., Li, X., and Ye, Y. PCR: proxy-based contrastive replay for online class-incremental continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition,*, pp. 24246–24255, 2023.

Liu, B., Xiao, X., and Stone, P. A lifelong learning approach to mobile robot navigation. *IEEE Robotics and Automation Letters*, 6(2):1090–1096, 2021.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020.

Liu, Y., Chen, T., and Yang, Q. A secure federated transfer learning framework. *IEEE Intelligent Systems*, 35(4): 70–82, 2018.

Liu, Y., Zhang, X., Kang, Y., Li, L., Chen, T., Hong, M., and Yang, Q. Fedbcd: A communication-efficient collaborative learning framework for distributed features. *IEEE Transactions on Signal Processing*, 70:4277–4290, 2022.

Liu, Y., Hong, X., Tao, X., Dong, S., Shi, J., and Gong, Y. Model behavior preserving for class-incremental learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):7529–7540, 2023.

Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y., and Yang, Q. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–20, 2024.

Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A. D., and Van De Weijer, J. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1273–1282, 2017.

Mesoudi, A. Cultural selection and biased transformation: two dynamics of cultural evolution. *Philosophical Transactions of the Royal Society B*, 376(1828):20200053, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8026–8037, 2019.

Pelosin, F., Jha, S., Torsello, A., Raducanu, B., and van de Weijer, J. Towards exemplar-free continual learning in vision transformers: an account of attention, functional and weight regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3820–3829, 2022.

Qi, W., Zhang, R., Zhou, J., Zhang, H., Xie, Y., and Jing, X. A resource-efficient cross-domain sensing method for device-free gesture recognition with federated transfer learning. *IEEE Transactions on Green Communications and Networking*, 7(1):393–400, 2023.

Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning*, volume 162, pp. 18250–18280. PMLR, 2022.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. icarl: Incremental classifier and representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Shi, Y., Liang, J., Zhang, W., Tan, V. Y. F., and Bai, S. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In *International Conference on Learning Representations*, 2023.

Thompson, B., Van Opheusden, B., Sumers, T., and Griffiths, T. Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science*, 376(6588):95–98, 2022.

Tiwari, R., Killamsetty, K., Iyer, R. K., and Shenoy, P. GCR: gradient coreset based replay buffer selection for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 99–108, 2022.

Tukey, J. W. et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

Wang, F.-Y., Zhou, D.-W., Ye, H.-J., and Zhan, D.-C. Foster: Feature boosting and compression for class-incremental learning. In *European Conference on Computer Vision*, pp. 398–414. Springer, 2022a.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623, 2020.

Wang, J., Chen, Y., Zheng, Z., Li, X., Cheng, M.-M., and Hou, Q. Crosskd: Cross-head knowledge distillation for dense object detection. *arXiv preprint arXiv:2306.11369*, 2023a.

Wang, Q.-F., Geng, X., Lin, S.-X., Xia, S.-Y., Qi, L., and Xu, N. Learngene: From open-world to your learning task. In *Association for the Advancement of Artificial Intelligence*, volume 36, pp. 8557–8565, 2022b.

Wang, Y., Ma, Z., Huang, Z., Wang, Y., Su, Z., and Hong, X. Isolation and impartial aggregation: A paradigm of incremental learning without interference. In *Association for the Advancement of Artificial Intelligence*, volume 37, pp. 10209–10217, 2023b.

Wu, W. and Zhang, Y. An efficient intrusion detection method using federated transfer learning and support vector machine with privacy-preserving. *Intelligent Data Analysis*, 27(4):1121–1141, 2023.

Yan, Q., Gong, D., Liu, Y., van den Hengel, A., and Shi, J. Q. Learning bayesian sparse networks with full experience replay for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 109–118, 2022.

Yan, S., Xie, J., and He, X. Der: Dynamically expandable representation for class incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2021.

Yang, B., Lin, M., Zhang, Y., Liu, B., Liang, X., Ji, R., and Ye, Q. Dynamic support network for few-shot class incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2945–2951, 2022.

Yang, S., Liu, L., and Xu, M. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations*, 2021.

Yoon, J., Jeong, W., Lee, G., Yang, E., and Hwang, S. J. Federated continual learning with weighted inter-client transfer. In *International Conference on Machine Learning*, volume 139, pp. 12073–12086. PMLR, 2021.

Zhang, C., Song, N., Lin, G., Zheng, Y., Pan, P., and Xu, Y. Few-shot incremental learning with continually evolved classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2021a.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., and Gao, Y. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021b.

Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S., and Wu, C. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, volume 162, pp. 26311–26329. PMLR, 2022.

Zhao, B., Xiao, X., Gan, G., Zhang, B., and Xia, S.-T. Maintaining discrimination and fairness in class incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13208–13217, 2020.

Zhou, D., Wang, F., Ye, H., Ma, L., Pu, S., and Zhan, D. Forward compatible few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition,CVPR*, pp. 9036–9046, 2022a.

Zhou, D.-W., Ye, H.-J., Ma, L., Xie, D., Pu, S., and Zhan, D.-C. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12816–12831, 2022b.

Zhou, D.-W., Wang, Q.-W., Ye, H.-J., and Zhan, D.-C. A model or 603 exemplars: Towards memory-efficient class-incremental learning. In *International Conference on Learning Representations*, 2023.

A. Supplementary Material

In the supplementary material, we provide more details about the implementations and experimental results mentioned in the main paper, as well as additional theoretical analysis and discussions. The supplementary material is organized as follows:

- In Section B, we provide a comprehensive overview of the compared methods employed in the main paper, as well as a detailed description of the datasets.
- In Section C, we report the full experimental results on two versions of the dataset split.
- In Section D, we provide the full theoretical analysis of sociability information and formalization of knowledge-driven knowledge distillation formula.

B. Implementation Details

In this section, we provide a detailed description of the compared methods and the datasets in the main paper. Our agent is deployed in PyTorch (Paszke et al., 2019) with an NVIDIA RTX 3090 GPU and trained with a batch size of 128 for 500 epochs, and we use SGD with momentum for optimization. The learning rate is set to 0.005, the energy anchor is set to -20, λ_1 is set to 1, and λ_2 is set to 0.1. All results are reported in means over 3 trials.

B.1. Compared Methods

In this subsection, we introduce the compared methods in the main paper. These methods are as follows:

- **DER (Yan et al., 2021):** classical continual learning (CL) method, which creates a novel backbone for each new incremental task. These individual backbones are concatenated to facilitate the learning of a unified classifier.
- **EWC (Kirkpatrick et al., 2017):** classical CL method, which attenuates the impact on essential parameters from prior tasks.
- **FOSTER (Wang et al., 2022a):** classical CL method, which builds upon DER by introducing a model compression stage to manage the memory budget.
- **iCaRL (Rebuffi et al., 2017):** classical CL method, which classifies based on a nearest-mean-of-exemplars rule and incorporates knowledge distillation and prototype rehearsal for representation learning.
- **LwF (Li & Hoiem, 2017):** classical CL method, which employs knowledge distillation as a regularization term to address the issue of forgetting. This regularization relies on the supervision of the old model.
- **MEMO (Zhou et al., 2023):** classical CL method, which generates new residual layers instead of an entire backbone to reduce memory costs.
- **PODNet (Douillard et al., 2020):** classical CL method, which applies a spatial-based distillation loss throughout the model, and a representation is comprised of multiple proxy vectors for each class.
- **WA (Zhao et al., 2020):** classical CL method, which employs knowledge distillation to preserve discrimination within old classes and weight aligning to correct biased weights in fully connected layers.
- **FedAvg (McMahan et al., 2017):** classical federated learning (FL) method, which employs an averaging scheme to amalgamate locally trained models into a global model.
- **FedAvgM (Hsu et al., 2019):** classical FL method, which explores the effect of non-identical data distributions on classification and designs the data synthesis method with a continuous range of identicalness.
- **FedNova (Wang et al., 2020):** classical FL method, which presents a comprehensive framework for analyzing the convergence behaviors of federated heterogeneous optimization algorithms.
- **FedDecorr (Shi et al., 2023):** classical FL method, which incorporates a regularization term into local training, fostering the decorrelation of distinct dimensions within representations.
- **FedProx (Li et al., 2020):** classical FL method, which solves heterogeneity in the federated network by a generalization and re-parametrization of FedAvg.
- **FedSAM (Qu et al., 2022):** classical FL method, which advances a momentum-based FL approach, bridging the local and global models through sharpness-aware minimization.
- **MOON (Li et al., 2021):** classical FL method, which leverages the inherent similarity within model representations to rectify local training deviations within individual nodes.
- **GLFC (Dong et al., 2022):** classical federated CL method, which incorporates class-aware gradient compensation loss and class-semantic relation distilling loss.

Datasets	Classes	Training instances	Testing instances	Detailed classes
CIFAR10-5-2	expert-class	10,000	2,000	airplane, automobile
	general-class-1	10,000	2,000	bird, cat
	general-class-2	10,000	2,000	deer, dog
	general-class-3	10,000	2,000	frog, horse
	general-class-4	10,000	2,000	ship, truck
CIFAR10-2-5	expert-class	25,000	5,000	airplane, automobile, bird, cat, deer
	general-class-1	25,000	5,000	dog, frog, horse, ship, truck
CIFAR100-4-25	expert-class	12,500	2,500	apple, aquarium_fish, baby, bear, beaver, bed, bee, beetle, bicycle, bottle, bowl, boy, bridge, bus, butterfly, camel, can, castle, caterpillar, cattle, chair, chimpanzee, clock, cloud, cockroach
	general-class-1	12,500	2,500	couch, crab, crocodile, cup, dinosaur, dolphin, elephant, flatfish, forest, fox, girl, hamster, house, kangaroo, keyboard, lamp, lawn_mower, leopard, lion, lizard, lobster, man, maple_tree, motorcycle, mountain
	general-class-2	12,500	2,500	mouse, mushroom, oak_tree, orange, orchid, otter, palm_tree, pear, pickup_truck, pine_tree, plain, plate, poppy, porcupine, possum, rabbit, raccoon, ray, road, rocket, rose, sea, seal, shark, shrew
	general-class-3	12,500	2,500	skunk, skyscraper, snail, snake, spider, squirrel, streetcar, sunflower, sweet_pepper, table, tank, telephone, television, tiger, tractor, train, trout, tulip, turtle, wardrobe, whale, willow_tree, wolf, woman, worm
CIFAR100-5-20	expert-class	10,000	2,000	apple, aquarium_fish, baby, bear, beaver, bed, bee, beetle, bicycle, bottle, bowl, boy, bridge, bus, butterfly, camel, can, castle, caterpillar, cattle
	general-class-1	10,000	2,000	chair, chimpanzee, clock, cloud, cockroach, couch, crab, crocodile, cup, dinosaur, dolphin, elephant, flatfish, forest, fox, girl, hamster, house, kangaroo, keyboard
	general-class-2	10,000	2,000	lamp, lawn_mower, leopard, lion, lizard, lobster, man, maple_tree, motorcycle, mountain, mouse, mushroom, oak_tree, orange, orchid, otter, palm_tree, pear, pickup_truck, pine_tree
	general-class-3	10,000	2,000	plain, plate, poppy, porcupine, possum, rabbit, raccoon, ray, road, rocket, rose, sea, seal, shark, shrew, skunk, skyscraper, snail, snake, spider
	general-class-4	10,000	2,000	squirrel, streetcar, sunflower, sweet_pepper, table, tank, telephone, television, tiger, tractor, train, trout, tulip, turtle, wardrobe, whale, willow_tree, wolf, woman, worm

Table 7: Detailed datasets.

B.2. Datasets

In this subsection, we provide an introduction to the datasets used in the main paper. We evaluate the performance on CIFAR10 (Krizhevsky et al., 2009) and CIFAR100 (Krizhevsky et al., 2009). CIFAR10 and CIFAR100 both consist of 60,000 images. CIFAR10 has 10 classes, while CIFAR100 has 100 classes. We employ two versions of dataset partitioning. In the first version, CIFAR10 is evenly distributed among 5 agents, each assigned 2 classes (abbreviated as expert-class, general-class-1, general-class-2, general-class-3, and general-class-4), denoted as CIFAR10-5-2. Similarly, CIFAR100 is evenly divided among 4 agents, each managing 25 classes (abbreviated as expert-class, general-class-1, general-class-2, and general-class-3), denoted as CIFAR100-4-25. In the second version, CIFAR10 is evenly split among 2 agents, each han-

dling 5 classes (CIFAR10-2-5), and CIFAR100 is evenly distributed among 5 agents, each dealing with 20 classes (CIFAR100-5-20). Taking one agent as an example, facing the 10 classes in CIFAR10-2-5, the 10 classes are paired off into groups of five, and the groups of classes that this agent is good at are named as expert-class, while the groups of classes that this agent is not good at are named as general-class-1. In the case of CL with only 1 agent, we substitute the number of agents with sessions. We list the details of the datasets in Table 7.

C. Full Experimental Results

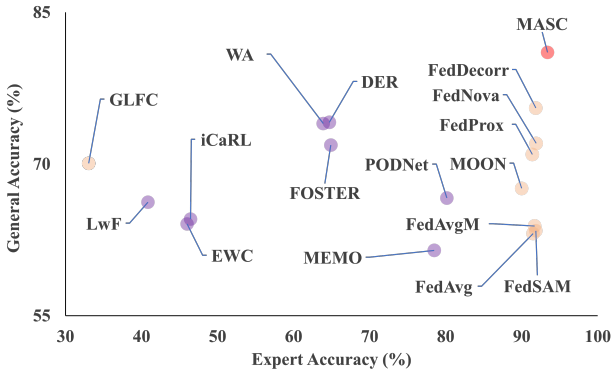
In this section, we report additional experimental results for the first version of the dataset split and comprehensive results for the second version of the dataset split.

C.1. The Experimental Results for the First Version of the Dataset Split

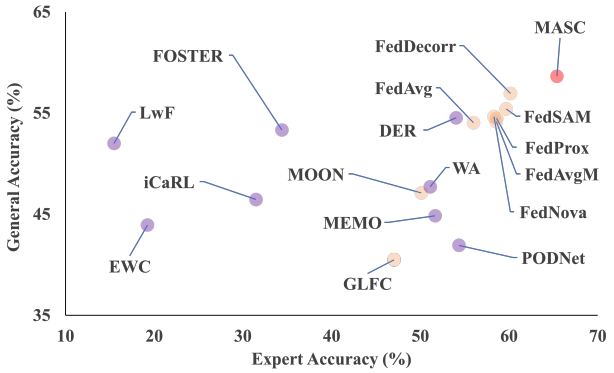
Analysis after growth: To provide a more intuitive comparison of the performance of different methods after growth on CIFAR10 and CIFAR100 datasets, we present a scatter plot. Closer proximity to the upper-right corner indicates better performance, as illustrated in Figure 5. It is evident that achieving high accuracy for both expert and general classes simultaneously poses a significant challenge. In this regard, Multi-Agent Socialized Collaboration (MASC) demonstrates promising performance.

Analysis of data-driven knowledge distillation: We leverage distinct datasets from individual teachers to calculate the respective mean and variance, facilitating the generation of synthetic data for agent training without direct access to real samples. To facilitate a more direct comparison of the performance of different methods on the generated CIFAR10 and CIFAR100 datasets before and after growth, we present the post-growth performance with a gray background. The performance trends in generated data, as shown in Table 8 and 9, align consistently with those observed in the original data. This suggests the meaningfulness of employing data-driven knowledge distillation, as it not only reduces data transmission costs but also enables effective agent training without accessing real samples.

Analysis of versatility and directionality: Versatility and directionality play pivotal roles in socialized learning (SL). The possession of versatility indicates the ability of agents to acquire knowledge from a broader spectrum of general classes, while directionality ensures the adept preservation of expert class performance. We conduct an analysis of versatility and directionality by juxtaposing the average accuracy across all classes and the accuracy differentials between expert classes and the best-performing general classes. As depicted in Figure 6, a notable pattern emerges where most



(a) Comparison on CIFAR10-5-2 dataset.



(b) Comparison on CIFAR100-4-25 dataset.

Figure 5: Performance comparison after growth. Different colors denote different learning paradigms.

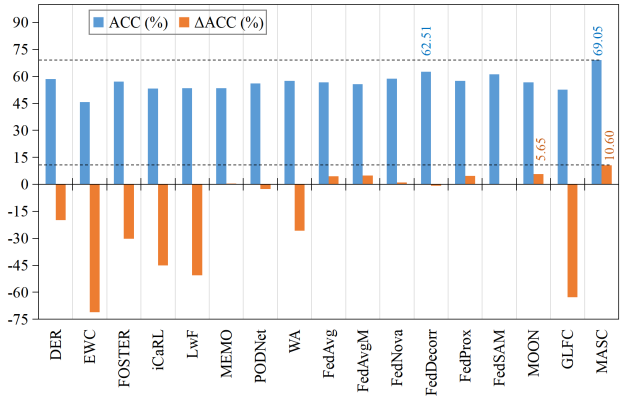
Method	Accuracy before growth			Accuracy after growth		
	Expert	General	Average	Expert	General	Average
DER (Yan et al., 2021)	89.81	0.00	44.91	59.35	58.23	58.79
EWC (Kirkpatrick et al., 2017)	89.32	0.00	44.66	18.55	52.45	35.50
FOSTER (Wang et al., 2022a)	89.79	0.00	44.90	53.45	57.95	55.70
iCaRL (Rebuffi et al., 2017)	89.74	0.00	44.87	40.85	56.43	48.64
LwF (Li & Hoiem, 2017)	89.58	0.00	44.79	32.40	58.84	45.62
MEMO (Zhou et al., 2023)	89.63	0.00	44.82	60.65	51.61	56.13
PODNet (Douillard et al., 2020)	89.69	0.00	44.85	65.50	53.74	59.62
WA (Zhao et al., 2020)	89.83	0.00	44.92	54.40	58.34	56.37
FedAvg (McMahan et al., 2017)	88.95	0.00	44.48	80.30	50.90	65.60
FedAvgM (Hsu et al., 2019)	88.58	0.00	44.29	76.75	50.44	63.59
FedNova (Wang et al., 2020)	89.09	0.00	44.55	80.20	53.48	66.84
FedDecorr (Shi et al., 2023)	89.15	0.00	44.58	78.60	58.49	68.54
FedProx (Li et al., 2020)	89.01	0.00	44.51	78.40	52.31	65.36
FedSAM (Qu et al., 2022)	89.11	0.00	44.56	**81.80**	56.04	**68.92**
MOON (Li et al., 2021)	88.65	0.00	44.33	79.65	50.89	65.27
GLFC (Dong et al., 2022)	89.61	0.00	44.81	25.90	**59.28**	42.59
MASC	89.40	0.00	44.70	**84.00**$_{(+2.2)}$	**65.31**$_{(+6.03)}$	**74.66**$_{(+5.74)}$

Table 8: Comparison of detailed accuracy across different classes before and after growth on generated CIFAR10-5-2 dataset. The 1^{st}/2^{nd} best results are indicated in red/blue.

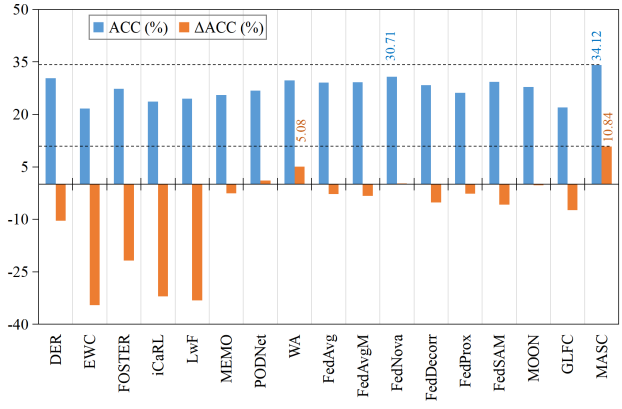
CL methods manifest catastrophic forgetting in expert class performance (indicated by negative values in orange bars). Simultaneously, the majority of FL methods struggle to sustain directionality in expert class performance, as evidenced by orange bar values approaching zero.

Method	Accuracy before growth			Accuracy after growth		
	Expert	General	Average	Expert	General	Average
DER (Yan et al., 2021)	44.28	0.00	22.14	29.12	30.69	29.91
EWC (Kirkpatrick et al., 2017)	44.04	0.00	22.02	10.16	25.53	17.85
FOSTER (Wang et al., 2022a)	44.19	0.00	22.10	16.84	**30.76**	23.80
iCaRL (Rebuffi et al., 2017)	44.07	0.00	22.04	13.32	27.04	20.18
LwF (Li & Hoiem, 2017)	44.13	0.00	22.07	11.56	28.87	20.21
MEMO (Zhou et al., 2023)	44.17	0.00	22.09	28.92	24.44	26.68
PODNet (Douillard et al., 2020)	44.11	0.00	22.06	33.20	24.64	28.92
WA (Zhao et al., 2020)	44.23	0.00	22.12	**37.32**	27.13	**32.23**
FedAvg (McMahan et al., 2017)	43.85	0.00	21.93	29.20	29.03	29.11
FedAvgM (Hsu et al., 2019)	43.88	0.00	21.94	26.96	30.00	28.48
FedNova (Wang et al., 2020)	43.96	0.00	21.98	32.88	29.99	31.43
FedDecorr (Shi et al., 2023)	43.82	0.00	21.91	25.28	29.44	27.36
FedProx (Li et al., 2020)	43.77	0.00	21.89	24.68	26.69	25.69
FedSAM (Qu et al., 2022)	43.91	0.00	21.96	26.36	30.25	28.31
MOON (Li et al., 2021)	43.79	0.00	21.90	29.56	27.32	28.44
GLFC (Dong et al., 2022)	44.08	0.00	22.04	27.74	20.02	23.88
MASC	44.16	0.00	22.08	**43.52**$_{(+6.2)}$	**30.99**$_{(+0.23)}$	**37.25**$_{(+5.02)}$

Table 9: Comparison of detailed accuracy across different classes before and after growth on generated CIFAR100-4-25 dataset. The 1^{st}/2^{nd} best results are indicated in red/blue.



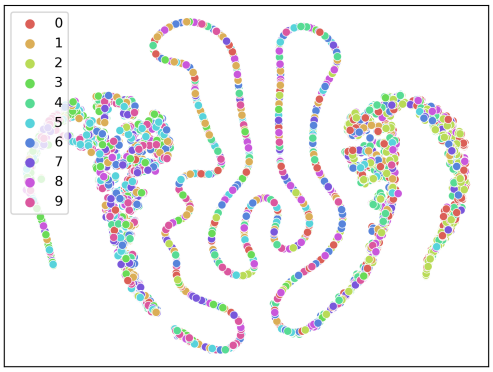
(a) Comparison on generated CIFAR10-5-2 dataset.



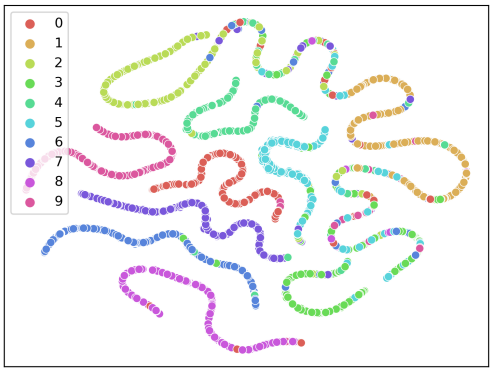
(b) Comparison on generated CIFAR100-4-25 dataset.

Figure 6: Analysis of versatility and directionality. Blue bars denote the average accuracy across all classes, while orange bars denote the difference in accuracy between expert classes and the best-performing general classes. Blue and orange bars correlate positively with versatility and directionality, respectively.

Analysis of t-SNE visualization: We visualize the representations of MASC before and after growth to investigate the effectiveness of SL. Observing Figure 7a, it is evident that before growth, agents struggle to cluster the same class and separate different classes, mainly due to the absence of exposure to general classes. In contrast, observing Figure 7b, after growth, MASC exhibited promising performance, demonstrating not only the ability to classify all classes but also more compact intra-class variations and clearer inter-class boundaries. This validates the effectiveness of SL.



(a) Before growth.

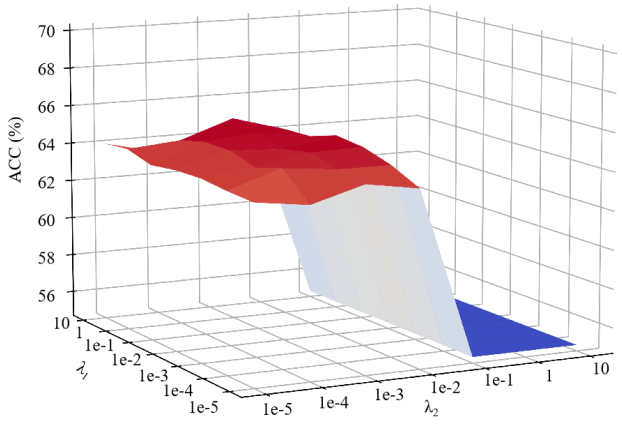


(b) After growth.

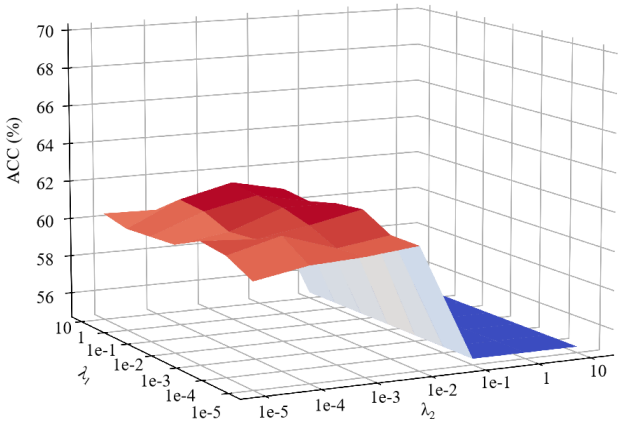
Figure 7: t-SNE visualization on CIFAR10-5-2 dataset before and after growth.

Analysis of trade-off hyperparameters: To assess the impact of λ_1 and λ_2, we varied their values within the range {1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10}. As illustrated in Figure 8, when λ_2 exceeds 1e-2, there is a significant performance decrease. This is attributed to the dominance of the alignment loss, making it challenging for the agent to

effectively train from ground truth and other agents. Regarding λ_1, our method demonstrates robustness. Through hyperparameter analysis experiments, setting λ_1 and λ_2 to 1 and 1e-3, respectively, yields superior performance for MASC.



(a) Accuracy of expert classes.



(b) Accuracy of all classes.

Figure 8: Parameter analysis on CIFAR100-4-25 dataset.

C.2. The Experimental Results for the Second Version of the Dataset Split

In the main paper, we elucidate the detailed process of training. For a clearer understanding of inference, we have described the algorithm in Algorithm 2.

As demonstrated in Tables 10, 11, 12, 13, 14, 15 and Figures 9, 10, consistent with the results from the first data split version, our proposed MASC exhibits commendable overall accuracy. It is evident that most CL methods show a noticeable trend of learning new knowledge while forgetting old knowledge. Although they perform well in general classes, this achievement comes at the cost of sacrificing expert class performance, contradicting our original intention. In summary, MASC provides valuable insights into SL,

Algorithm 2 Inference for MASC.

Given components: Backbone $f_b^s(\cdot)$, student classifier $f_c^s(\cdot)$, teacher classifier $f_c^{t_i'}(\cdot)$;
Input: Test sample $\boldsymbol{x}$;
Output: Final prediction y^*;

1: Calculate image feature $f_b^s(\boldsymbol{x})$;
2: Calculate the energy of the student $-\mathcal{F}^s(\boldsymbol{x})$;
3: Calculate the energy of the grown teacher $-\mathcal{F}^{t_i'}(\boldsymbol{x})$;
4: Obtain the classifier with the highest energy k^*;
5: Return final prediction y^*.

showcasing a high degree of flexibility that allows dynamic adjustments based on the actual number of agents.

Method	Accuracy before growth			Accuracy after growth		
	Expert	General	Average	Expert	General	Average
DER (Yan et al., 2021)	87.86	0.00	43.93	64.02	92.72	78.37
EWC (Kirkpatrick et al., 2017)	87.93	0.00	43.97	29.80	93.70	61.75
FOSTER (Wang et al., 2022a)	87.96	0.00	43.98	70.64	92.06	**81.35**
iCaRL (Rebuffi et al., 2017)	87.83	0.00	43.92	45.16	93.12	69.14
LwF (Li & Hoiem, 2017)	87.82	0.00	43.91	45.24	89.86	67.55
MEMO (Zhou et al., 2023)	87.77	0.00	43.89	76.78	69.72	73.25
PODNet (Douillard et al., 2020)	87.79	0.00	43.90	79.48	82.64	81.06
WA (Zhao et al., 2020)	87.91	0.00	43.96	71.68	90.48	81.08
FedAvg (McMahan et al., 2017)	87.72	0.00	43.86	76.38	78.30	77.34
FedAvgM (Hsu et al., 2019)	87.66	0.00	43.83	76.36	76.36	76.36
FedNova (Wang et al., 2020)	87.73	0.00	43.87	77.36	78.64	78.00
FedDecorr (Shi et al., 2023)	87.82	0.00	43.91	**81.56**	79.72	80.64
FedProx (Li et al., 2020)	87.76	0.00	43.88	80.74	80.42	80.58
FedSAM (Qu et al., 2022)	87.69	0.00	43.85	78.18	76.06	77.12
MOON (Li et al., 2021)	87.62	0.00	43.81	72.82	72.40	72.61
GLFC (Dong et al., 2022)	87.68	0.00	43.84	54.06	**94.10**	74.08
MASC	87.80	0.00	43.90	**86.68**$_{(+5.12)}$	**78.80**	**82.74**$_{(+1.39)}$

Table 10: Comparison of detailed accuracy across different classes before and after growth on CIFAR10-2-5 dataset. The 1^{st}/2^{nd} best results are indicated in red/blue.

Method	Accuracy before growth			Accuracy after growth		
	Expert	General	Average	Expert	General	Average
DER (Yan et al., 2021)	67.81	0.00	33.91	46.30	53.69	49.99
EWC (Kirkpatrick et al., 2017)	67.75	0.00	33.88	13.93	37.03	25.48
FOSTER (Wang et al., 2022a)	67.66	0.00	33.83	34.80	43.39	39.09
iCaRL (Rebuffi et al., 2017)	67.29	0.00	33.65	20.35	40.66	30.51
LwF (Li & Hoiem, 2017)	67.12	0.00	33.56	15.73	37.88	26.80
MEMO (Zhou et al., 2023)	67.62	0.00	33.81	44.50	43.91	44.21
PODNet (Douillard et al., 2020)	67.43	0.00	33.72	51.65	42.73	47.19
WA (Zhao et al., 2020)	67.51	0.00	33.76	50.80	47.75	49.28
FedAvg (McMahan et al., 2017)	67.02	0.00	33.51	51.10	54.89	52.99
FedAvgM (Hsu et al., 2019)	67.05	0.00	33.53	50.35	55.76	53.06
FedNova (Wang et al., 2020)	67.13	0.00	33.57	48.80	56.44	52.62
FedDecorr (Shi et al., 2023)	67.29	0.00	33.65	**53.25**	**57.53**	**55.39**
FedProx (Li et al., 2020)	67.09	0.00	33.55	51.45	55.70	53.58
FedSAM (Qu et al., 2022)	67.16	0.00	33.58	51.25	56.46	53.86
MOON (Li et al., 2021)	66.98	0.00	33.49	44.20	48.49	46.34
GLFC (Dong et al., 2022)	67.76	0.00	33.88	14.96	48.43	31.69
MASC	67.40	0.00	33.70	**65.95**$_{(+12.7)}$	**58.09**$_{(+0.56)}$	**62.02**$_{(+6.63)}$

Table 11: Comparison of detailed accuracy across different classes before and after growth on CIFAR100-5-20 dataset. The 1^{st}/2^{nd} best results are indicated in red/blue.

D. Theoretical Analysis

Below, we will elucidate the omitted proofs in Section 3 and the formalized formulas for data-driven knowledge distillation presented in Section 5 (see main paper).

Method	Accuracy before growth			Accuracy after growth		
	Expert	General	Average	Expert	General	Average
DER (Yan et al., 2021)	72.68	0.00	36.34	54.18	72.56	**63.37**
EWC (Kirkpatrick et al., 2017)	72.66	0.00	36.33	15.58	80.08	47.83
FOSTER (Wang et al., 2022a)	72.59	0.00	36.30	46.36	77.44	61.90
iCaRL (Rebuffi et al., 2017)	72.51	0.00	36.26	34.42	78.10	56.26
LwF (Li & Hoiem, 2017)	72.55	0.00	36.28	31.30	76.66	53.98
MEMO (Zhou et al., 2023)	72.47	0.00	36.24	54.08	60.64	57.36
PODNet (Douillard et al., 2020)	72.49	0.00	36.25	46.60	71.44	59.02
WA (Zhao et al., 2020)	72.62	0.00	36.31	53.58	73.00	63.29
FedAvg (McMahan et al., 2017)	72.34	0.00	36.17	48.32	66.92	57.62
FedAvgM (Hsu et al., 2019)	72.41	0.00	36.21	52.90	68.70	60.80
FedNova (Wang et al., 2020)	72.45	0.00	36.23	**64.22**	59.28	61.75
FedDecorr (Shi et al., 2023)	72.49	0.00	36.25	62.78	63.84	63.31
FedProx (Li et al., 2020)	72.37	0.00	36.19	56.62	63.70	60.16
FedSAM (Qu et al., 2022)	72.43	0.00	36.22	52.18	70.68	61.43
MOON (Li et al., 2021)	72.31	0.00	36.16	46.50	67.32	56.91
GLFC (Dong et al., 2022)	72.46	0.00	36.23	41.72	**80.66**	61.19
MASC	72.52	0.00	36.26	**69.24**$_{(+5.02)}$	**62.16**	**65.70**$_{(+2.33)}$

Table 12: Comparison of detailed accuracy across different classes before and after growth on generated CIFAR10-2-5 dataset. The 1^{st}/2^{nd} best results are indicated in red/blue.

Method	Accuracy before growth			Accuracy after growth		
	Expert	General	Average	Expert	General	Average
DER (Yan et al., 2021)	41.42	0.00	20.71	26.10	**30.16**	28.13
EWC (Kirkpatrick et al., 2017)	41.33	0.00	20.67	2.48	20.33	11.40
FOSTER (Wang et al., 2022a)	41.27	0.00	20.64	**36.05**	18.81	27.43
iCaRL (Rebuffi et al., 2017)	41.19	0.00	20.60	8.15	23.08	15.61
LwF (Li & Hoiem, 2017)	41.23	0.00	20.62	6.93	20.76	13.85
MEMO (Zhou et al., 2023)	41.36	0.00	20.68	28.40	23.03	25.71
PODNet (Douillard et al., 2020)	41.39	0.00	20.70	31.60	25.24	28.42
WA (Zhao et al., 2020)	41.53	0.00	20.77	34.35	28.16	**31.26**
FedAvg (McMahan et al., 2017)	41.09	0.00	20.55	24.70	27.35	26.03
FedAvgM (Hsu et al., 2019)	41.19	0.00	20.60	24.30	29.31	26.81
FedNova (Wang et al., 2020)	41.22	0.00	20.61	24.90	29.71	27.31
FedDecorr (Shi et al., 2023)	41.12	0.00	20.56	24.75	27.74	26.24
FedProx (Li et al., 2020)	41.05	0.00	20.53	20.45	24.54	22.49
FedSAM (Qu et al., 2022)	41.17	0.00	20.59	23.60	28.46	26.03
MOON (Li et al., 2021)	41.13	0.00	20.57	21.70	28.79	25.24
GLFC (Dong et al., 2022)	41.31	0.00	20.66	6.55	25.64	16.09
MASC	41.25	0.00	20.63	**39.00**$_{(+2.95)}$	**30.90**$_{(+0.74)}$	**34.95**$_{(+3.69)}$

Table 13: Comparison of detailed accuracy across different classes before and after growth on generated CIFAR100-5-20 dataset. The 1^{st}/2^{nd} best results are indicated in red/blue.

Dataset	Method	Accuracy before growth			Accuracy after growth		
		Expert	General	Average	Expert	General	Average
CIFAR10	Agent-1	87.80	0.00	43.90	86.68	78.80	**82.74**
	Agent-2	92.24	0.00	46.12	89.26	78.62	**83.94**
CIFAR100	Agent-1	67.40	0.00	33.70	65.95	58.09	**62.02**
	Agent-2	68.70	0.00	34.35	65.10	58.88	**61.99**
	Agent-3	70.30	0.00	35.15	65.30	58.06	**61.68**
	Agent-4	66.70	0.00	33.35	63.15	57.81	**60.48**
	Agent-5	69.85	0.00	34.93	64.30	60.83	**62.56**

Table 14: Comparison of different agents before and after growth on CIFAR10-2-5 and CIFAR100-5-20.

D.1. Proof of Theorem 3.4

Proof. We utilize the previous results (Cover, 1999; Feder & Merhav, 1994; Li et al., 2023b):

$$-\log\left(1-P_{e_c}^{mul}\right) \leq H(Y \mid X_{\mathcal{A}_1}, X_{\mathcal{A}_2}), \tag{18}$$

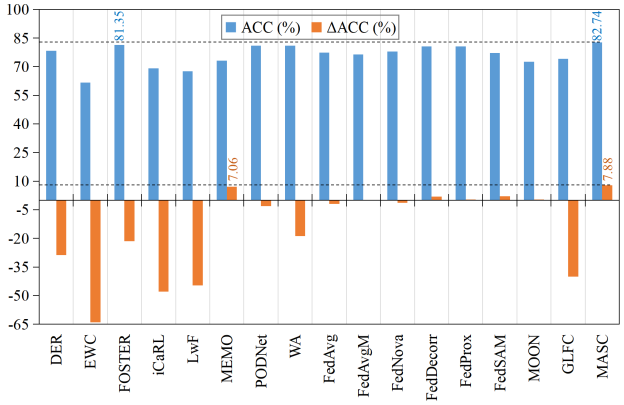
$$H(Y \mid X_{\mathcal{A}_1}, X_{\mathcal{A}_2}) \leq \log 2 + P_{e_c}^{mul} \log |Y|. \tag{19}$$

Combine the two inequalities and put $P_{e_c}^{mul}$ in the middle:

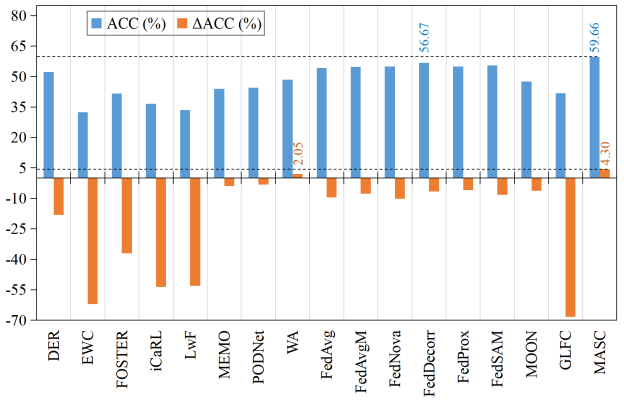
$$\frac{H(Y|X_{\mathcal{A}_1},X_{\mathcal{A}_2})-\log 2}{\log|Y|} \leq P_{e_c}^{mul} \leq 1-\exp(-H(Y \mid X_{\mathcal{A}_1}, X_{\mathcal{A}_2})), \tag{20}$$

Dataset	Method	Accuracy before growth			Accuracy after growth		
		Expert	General	Average	Expert	General	Average
CIFAR10	Agent-1	72.52	0.00	36.26	69.24	62.16	**65.70**
	Agent-2	79.66	0.00	39.83	73.90	66.08	**69.99**
CIFAR100	Agent-1	41.25	0.00	20.63	39.00	30.90	**34.95**
	Agent-2	42.85	0.00	21.43	39.40	30.65	**35.03**
	Agent-3	43.30	0.00	21.65	38.30	33.19	**35.74**
	Agent-4	41.70	0.00	20.85	37.05	33.44	**35.24**
	Agent-5	44.50	0.00	22.25	41.90	31.01	**36.46**

Table 15: Comparison of different agents before and after growth on generated CIFAR10-2-5 and CIFAR100-5-20.



(a) Comparison on CIFAR10-2-5 dataset.

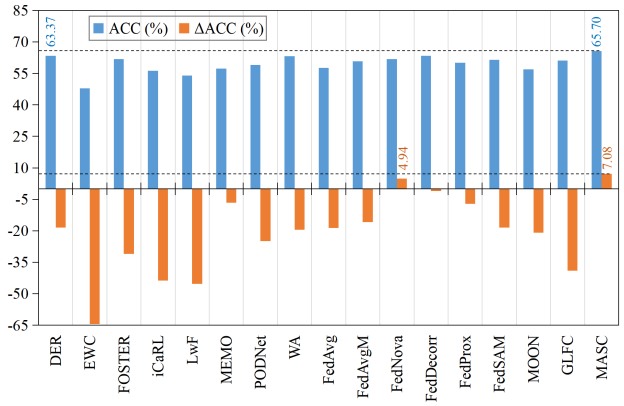


(b) Comparison on CIFAR100-5-20 dataset.

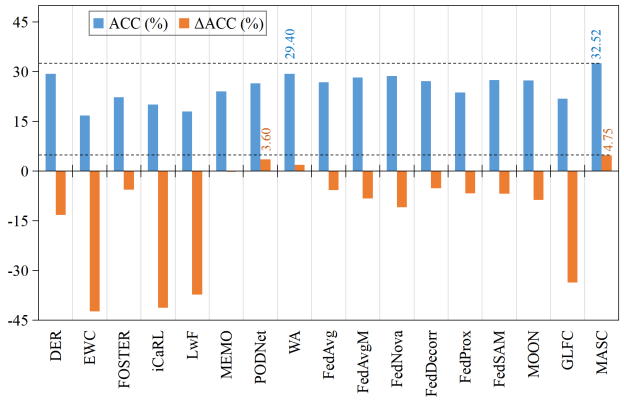
Figure 9: Analysis of versatility and directionality on original datasets. Blue bars denote the average accuracy across all classes, while orange bars denote the difference in accuracy between expert classes and the best-performing general classes. Blue and orange bars correlate positively with versatility and directionality, respectively.

which is the first result in the theorem. Then we apply the results to $P_{e_c}^{sin}$:

$$\frac{H(Y|X_{\mathcal{A}_1})-\log 2}{\log|Y|} \leq P_{e_c}^{sin} \leq 1-\exp(-H(Y \mid X_{\mathcal{A}_1})). \tag{21}$$



(a) Comparison on generated CIFAR10-2-5 dataset.



(b) Comparison on generated CIFAR100-5-20 dataset.

Figure 10: Analysis of versatility and directionality on generated datasets. Blue bars denote the average accuracy across all classes, while orange bars denote the difference in accuracy between expert classes and the best-performing general classes. Blue and orange bars correlate positively with versatility and directionality, respectively.

Since

$$\begin{aligned}\Phi_{X_{\mathcal{A}_2}} &= I(X_{\mathcal{A}_2}; Y \mid X_{\mathcal{A}_1}) \\ &= I(Y; X_{\mathcal{A}_1}, X_{\mathcal{A}_2}) - I(Y; X_{\mathcal{A}_1}) \\ &= [H(Y) - H(Y \mid X_{\mathcal{A}_1}, X_{\mathcal{A}_2})] - [H(Y) - H(Y \mid X_{\mathcal{A}_1})] \\ &= H(Y|X_{\mathcal{A}_1}) - H(Y|X_{\mathcal{A}_1}, X_{\mathcal{A}_2}).\end{aligned} \tag{22}$$

We can derive

$$\frac{H(Y|X_{\mathcal{A}_1}) - \Phi_{X_{\mathcal{A}_2}} - \log 2}{\log |Y|} \leq P_{e_c}^{mul} \leq 1 - \exp(-H(Y \mid X_{\mathcal{A}_1}) + \Phi_{X_{\mathcal{A}_2}}). \tag{23}$$

□

D.2. Data-driven Knowledge Distillation

Inspired by (Yang et al., 2021), we posit that the feature distribution for expert classes follows Gaussian distribution, and aim to enhance this Gaussian resemblance across distributions. To achieve this, we initially apply Tukey's Ladder of Powers transformation to the features of the target task, as detailed in (Tukey et al., 1977). This transformation, belonging to a family of power transformations, is designed to reduce distribution skewness and render them more Gaussian-like. It operates as follows:

$$\boldsymbol{x} = \begin{cases} \tilde{\boldsymbol{x}}^{\lambda} & \text{if } \lambda \neq 0 \\ \log(\tilde{\boldsymbol{x}}) & \text{if } \lambda = 0 \end{cases}, \tag{24}$$

where $\tilde{\boldsymbol{x}}$ is a feature vector from an expert class, and λ is a hyper-parameter that adjusts the correction of the distribution. Setting λ to 1 recovers the original feature. Decreasing λ reduces the positive skew of the distribution, while increasing λ enhances it.

The mean of the feature vector from an expert class i is calculated as:

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n_i} \boldsymbol{x}_j}{n_i}, \tag{25}$$

where $\boldsymbol{x}_j$ is a feature vector of the j-th sample from the expert class i and n_i is the total number of samples in class i. Given the multidimensional nature of the feature vector $\boldsymbol{x}_j$, we utilize covariance to more accurately represent the variance between any two elements within the feature vector. The covariance matrix $\boldsymbol{\Sigma}_i$ for features from class i is calculated in the following manner:

$$\boldsymbol{\Sigma}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T. \tag{26}$$

To effectively utilize these transformed features in a target task, we compile a set of calibrated statistics $\mathbb{S}_y = \{(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)\}$ for each class y. We then generate a series of feature vectors labeled y by sampling from these calibrated Gaussian distributions, thereby aligning more closely with the assumed Gaussian feature distribution of expert classes, which is defined as:

$$\mathbb{D}_y = \{(\boldsymbol{x}, y) \mid \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \forall (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{S}^y\}. \tag{27}$$