Pattern Recognition 156 (2024) 110824

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Federated zero-shot learning with mid-level semantic knowledge transfer

Shitong Sun [a,*], Chenyang Si [b], Guile Wu [c], Shaogang Gong [a]

[a] Queen Mary University of London, London, E1 4NS, United Kingdom
[b] Nanyang Technological University, Singapore, 639798, Singapore
[c] Independent Researcher

ARTICLE INFO

Keywords:
Federated learning
Knowledge transfer

ABSTRACT

Conventional centralized deep learning paradigms are not feasible when data from different sources cannot be shared due to data privacy or transmission limitation. To resolve this problem, federated learning has been introduced to transfer knowledge across multiple sources (clients) with non-shared data while optimizing a globally generalized central model (server). Existing federated learning paradigms mostly focus on transmitting image encoders that take instance-sensitive images as input, making them less generalizable and vulnerable to privacy inference attacks. In contrast, in this work, we consider transferring mid-level semantic knowledge (such as attribute) which is not sensitive to specific objects of interest and therefore is more privacy-preserving and general. To this end, we formulate a new Federated Zero-Shot Learning (FZSL) paradigm to learn mid-level semantic knowledge at multiple local clients with non-shared local data and cumulatively aggregate a globally generalized central model for deployment. To improve model discriminative ability, we explore semantic knowledge available from either a language or a vision-language foundation model in order to enrich the mid-level semantic space in FZSL. Extensive experiments on five zero-shot learning benchmark datasets validate the effectiveness of our approach for optimizing a generalizable federated learning model with mid-level semantic knowledge transfer.

1. Introduction

Deep learning has gained great success in computer vision and natural language processing, but conventional deep learning paradigms mostly follow a centralized learning manner where data from different sources are collected to create a central database for model learning. With an increasing awareness of data privacy, decentralized deep learning becomes more desirable. To this end, federated learning [1] has been recently introduced to optimize local models (clients) with non-shared local data while learning a global generalized central model (server) by transferring knowledge across the clients and the server. This enables to protect data privacy inherently as local data are only used for training local models and only model parameters are transmitted across the clients and server. There have been a variety of federated learning methods for computer vision applications, such as person re-identification [2], action recognition [3] and medical image analysis [4–6].

One of the core challenges in federated learning in the real world is the statistical heterogeneity: the local data in different clients can be non-independent identically distributed (Non-IID), resulting in inconsistent update optimization directions of participant clients [7]. This statistical heterogeneity mainly includes differences in the client data concerning label space, feature space and data quantity. To address the challenges posed by data heterogeneity, early methods align the optimization direction among clients either through regularization terms [8,9] or by using an external dataset available to all clients [10]. Another line of work, personalized federated learning, solves this data heterogeneity problem by learning domain-specific knowledge for each client while benefiting from collaborative training [11]. However, both of these two lines of work can only extract in-distribution knowledge from participating clients without the ability to generalize to the open domain, making it challenging for a new potential client to join in. To deal with this concern, recent works extract generalizable knowledge from heterogeneous client data [12] and apply it to predict on unseen data. Although these works consider the generalization ability in federated learning, privacy protection against attacks is overlooked. As existing methods mostly transmit image encoders that take instance-sensitive image as input, rendering the system vulnerable to privacy inference attacks [13]. Attacker can reconstruct the model input by analyzing transmitted gradients. Once the current methods are attacked, the raw image can be recovered pixelwise. For visualization and a detailed discussion, refer to Fig. 4 and Section 4.2.

* Corresponding author.
E-mail address: shitong.sun@qmul.ac.uk (S. Sun).

https://doi.org/10.1016/j.patcog.2024.110824
Received 2 January 2024; Received in revised form 29 April 2024; Accepted 22 July 2024
Available online 29 July 2024

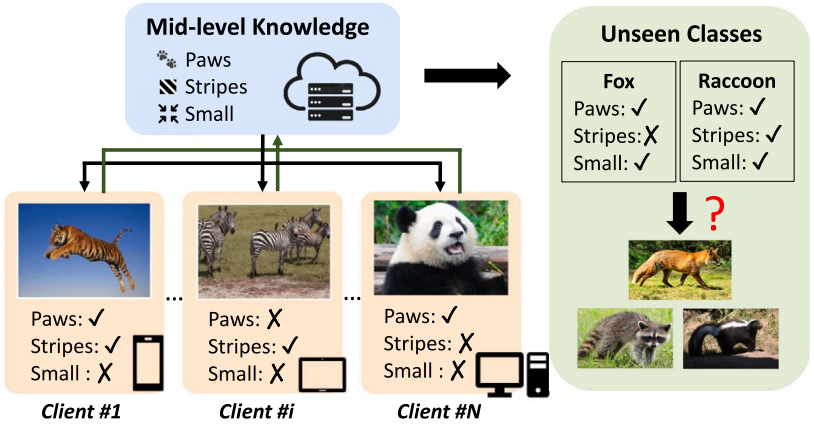


Fig. 1. An overview of federated zero-shot learning with mid-level semantic knowledge transfer. Each local client optimizes a local model with non-shared local data whilst a central server aggregates a global model by aggregating local model parameters. The server model will further be tested on unseen novel classes.

To mitigate the privacy leakage against reconstruction attacks, we draw inspiration from data anonymization methods in healthcare [14], where the organization has the requirement to publish data but keep the unique patient identity private. One solution is releasing quasi-identifiers like age, gender rather than personal identifiers like names. This inspires us a more secure approach: instead of transmitting models that take instance-sensitive images as input, transmitting models that take instance-insensitive attributes as input can help mitigate the risk of raw data leakage from the source. Moreover, extracting mid-level knowledge also boosts generalization to unseen data. As we can represent various classes using the same attribute dictionary, which is hypothesized in compositional learning: the number of attributes is finite, while the number of classes can be infinite [15]. This attribute knowledge transfer is also similar to human cognitive process [16]. In this paper, we refer to the conventional features encoding the instance-sensitive image feature as high-level features, and the features encoding attributes as mid-level features. Therefore, in federated learning with data heterogeneity, particularly for clients with different label spaces but a sharable attribute space, learning mid-level semantic knowledge transfer offers a new solution to protect privacy against reconstruction attacks while enhancing model generalization.

Limited works have explored mid-level knowledge transfer in decentralized learning scenarios. In contrast, in centralized learning scenarios, zero-shot learning (ZSL) is a well-established paradigm for learning mid-level knowledge. ZSL aims to learn a mid-level semantic mapping between image features and text labels (typically attributes) using seen object categories and then transfers knowledge to recognize unseen object categories with the help of the composition of shared attributes between seen and unseen categories. Existing ZSL methods [17,18] mostly focus on centralized learning scenarios, where training data from different label spaces is shared with a central data collection. Consequently, learning mid-level knowledge in ZSL under a decentralized learning paradigm remains an open question.

In this work, we formulate a new Federated Zero-Shot Learning (FZSL) paradigm, which aims to learn mid-level semantic knowledge for zero-shot learning across decentralized clients with heterogeneous data. An overview of FZSL is depicted in Fig. 1. Specifically, we consider there are multiple local clients where each client has an independent non-overlapping class label space whilst all clients share a common mid-level attribute space. This is an extremely non-IID data partition with a high degree of label distribution skewness and mean Kolmogorov–Smirnov value of one [19]. Then, we optimize local models (clients) with non-shared local data and learn a central generalized model (server) by transferring knowledges (model parameters) between the clients and the server. Further, the server model is tested on unseen novel classes utilizing the learned mid-level knowledge. With this paradigm, we innovatively learn mid-level semantic knowledge in a decentralized learning manner with data privacy protection by bridging federated learning and zero-shot learning. It cumulatively optimizes a generic mid-level attribute space from non-sharable distributed local data of different object categories. Instead of aggregating holistic models like traditional federated learning [1] or separating domain-specific classifiers like recent decentralized learning [20,21], we only aggregate generators across the clients and the server while discriminators are retained locally. This facilitates to learn more generalized knowledge and reduce the number of model parameters for communicating. Furthermore, to improve model discriminative ability, we utilize external knowledge by employing off-the-shelf foundation model (language model (e.g., RoBERTa [22]) or text encoder of vision-language model (e.g., CLIP [23])) to explore semantic knowledge augmentation to enrich the mid-level semantic space in FZSL. The text encoder is frozen in local clients and decoupled from the client–server aggregation process. With the help of the pre-trained text encoder supplying richer knowledge space, this semantic knowledge augmentation allows to learn a more generic knowledge to encode sample diversity as well as model scalability.

Our **contributions** are: (1) We propose to exploit mid-level semantic knowledge transfer for federated learning from independent non-overlapping class label spaces and introduce a new Federated Zero-Shot Learning paradigm. This paradigm illuminates a novel approach to enhance generalization and protect privacy against reconstruction attacks in federated learning scenarios involving heterogeneous data distributions, particularly for clients with different label spaces but a shared attribute space. (2) We formulate a baseline model, from which we further explore semantic knowledge augmentation by using either a language or a vision-language foundation model as external knowledge to learn a richer mid-level semantic space in FZSL. (3) We conduct extensive experiments on five zero-shot learning benchmark datasets to demonstrate the capability of our approach in learning a generalized federated learning model through mid-level semantic knowledge transfer.

2. Related work

2.1. Federated learning

Federated learning [1,9] is a recently introduced model learning paradigm aiming to learn a central model (server) with the collaboration of multiple local models (clients) under data privacy protection. It has been explored in various computer vision tasks, such as medical image analysis [4], person re-identification [21] and action recognition [3]. Conventional federated learning approaches, e.g., FedAvg [1], learn a sharable central model by aggregating holistic model parameters among different local models. These conventional methods

encounter challenges of statistical heterogeneity in real-word scenarios, where distribution of label space or feature space varies across local clients. To deal with label distribution skewness, where each client has access to a partial set of the whole class set, FedRS [24] selectively updates classifier weights of local clients for only the observed classes. FedLC [25] proposes a fine-grained calibrated loss to mitigate the deviation of local gradient updates. However, these methods encounter overlapping label spaces among participating clients and primarily in-distribution knowledge, which limits their ability to generalize to the open domain. In contrast, our work delves into the extreme scenario of label distribution shift, where there are non-overlapping label spaces among clients. To tackle this challenge, we introduce zero-shot learning with mid-level semantic knowledge transfer for federated learning, aiming to transfer generalized knowledge across heterogeneous clients.

Although there have been several seemingly related federated zero-shot learning studies [26–28], they mostly use a different definition and setting of *zero-shot* and none of these methods are aimed at bridging the gap between seen and unseen classes by learning mid-level semantic knowledge. ZSDG [26] implements zero-shot augmentation to generate existing categories and study a sharing class space. FedZKT [27] is based on zero-shot knowledge distillation with the purpose of transferring knowledge between clients and server without additional datasets. While our work is based on traditional zero shot learning task aimed to learn transferrable *mid-level* semantic attributes. Further, our FZSL is generalizable on *unseen* classes, while FedZKT and ZSDG are only tested on seen classes; our FZSL is learning from multiple independent *non-overlapping* class label spaces, while ZSDG [26] is studying a sharing class space. Among all, FedZSL [28] is the most similar to us, but it is based on high-level knowledge like other methods [26,27], while our FZSL learns to transfer mid-level semantic knowledge.

2.2. Zero shot learning

Zero shot learning (ZSL) aims to recognize unseen object categories leveraging seen categories for learning consistent semantic information to bridge seen and unseen categories. Current ZSL methods can broadly be divided into embedding based methods [29] and generative based methods [30]. Embedding based methods transfer from a visual space to a semantic space and classify unseen categories based on semantic similarity without any training data. In contrast, generative based methods learn a projection from a semantic space to a visual space, which enables to turn the zero shot learning task to a pseudo feature supervised learning task, alleviating overfitting [30]. Existing ZSL methods are following a centralized learning manner, while our work proposes a new federated zero-shot learning paradigm to transfer mid-level knowledge across different non-overlapping class label spaces with data privacy protection.

To handle instances in the testing stage with class labels not present during training, related topics include open-set recognition [31,32] and novel class discovery [33]. Open-set recognition aims to identify previously unseen classes and classify the seen classes. CGDL [31] employs variational auto-encoder with an additional classifier to categorize seen classes and detector to identity unseen classes. PROSER [32] introduces a learnable unseen class during training to address the challenge of unseen class diversity. However, these open-set recognition methods can only detect the unseen class without the ability of classify them. On the other hand, novel class discovery focuses on leveraging labeled data to classify unlabeled instances. Zhong et al. [33] employs contrastive learning by leveraging labeled samples to generate hard negative samples and classify unlabeled data. However, both of these tasks utilize holistic instance-sensitive images as input and would be vulnerable to attacks [13] if implemented in federated learning scenarios requiring privacy protection. In contrast, zero-shot learning enables the transfer of instance-insensitive attributes from seen to unseen classes, fostering generalized knowledge transfer through compositional learning. This approach not only facilitates classification of unseen data but also inherently provides privacy protection, a crucial requirement in federated learning scenarios.

2.3. Foundation models

Foundation models refer to models trained with a vast quantity of data and can be further adapted for various downstream tasks, such as BERT [34], RoBERTa [22], CLIP [23], etc. These models are usually learned by self-learning on unlabeled data and can predict underlying properties such as attributes, so they are scalable and potentially more useful than models trained on a limited label space. A concurrent work, PromptFL [35], has recently adapted a foundation model in federated learning to explore the potential representational power of the pretrained richer knowledge. However, it adapts the whole vision-language model (both image and text encoder) for model learning and focuses on general image classification. Critically, the image encoder of the foundation models cannot be used directly in zero shot learning image classification [16]. This is because that the ZSL hypothesis of disjoint training and test class labels [16,36] is not guaranteed in vision-language foundation models. To overcome this problem, we employ either a language model or a text encoder of a vision-language foundation model to augment *only* the mid-level semantic space in FZSL. And we use the ImageNet as the pretrained dataset and ensure non-overlapping of training and test classes guaranteed by [36].

3. Methodology

3.1. Problem definition

In this work, we study Federated Zero-Shot Learning (FZSL), where each client contains an independent non-overlapping class label space with non-shared local data. Suppose there are N local clients, where the ith client contains a training set $\mathcal{S}_i = \{\boldsymbol{x}, y\}$, with $y \in \mathcal{Y}_i$, subject to $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset, \forall i, j$, $\mathcal{Y}_1 \cup \mathcal{Y}_2 \cup \cdots \cup \mathcal{Y}_N = \mathcal{Y}_s$. $\mathcal{Y}_s$ denotes the seen class set. Each class can be characterized by a series of attributes within a mid-level attribute space, which is shared among both seen classes across all clients and unseen classes. Namely, each class y can be described by an attribute vector $\boldsymbol{a} = [a_1, a_2 \dots a_m]$ consisting of m individual concrete attributes. For instance, a_i represents the degree of the attribute 'stripes'. The objective of FZSL is to learn mid-level semantic knowledge across clients, and predict the unseen classes only based on their associated attributes. Namely, given $\boldsymbol{a}$ for class y, $y \in \mathcal{Y}_u$, $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$, we aim to construct a classifier $F : \mathcal{X} \rightarrow \mathcal{Y}$ for $\mathcal{Y}_u \subset \mathcal{Y}$, where $\mathcal{Y}_u$ is the unseen set.

3.2. Mid-level semantic knowledge transfer

To learn mid-level semantic knowledge transfer for federated learning, we formulate a baseline model which unifies federated learning and zero-shot learning in a decentralized learning paradigm. Since generative based zero-shot learning is capable of generating pseudo image features according to a consistent and generic mid-level attribute space, in this work, we employ a representative ZSL method, i.e., f-CLSWGAN [30], as the backbone. As shown in Fig. 2, the learning process consists of three iterative steps, namely local model learning, central model aggregation and local model reinitialization with the central model.

Training Stage. In each local client, with the non-shared local data $\mathcal{S}_i = \{\boldsymbol{x}, y\}$, the model learning process follows f-CLSWGAN [30]. A generator $G(\boldsymbol{z}, \boldsymbol{a}_g)$ learns to generate a CNN feature $\tilde{\boldsymbol{x}}$ in the input feature space $\mathcal{X}$ from random noise $\boldsymbol{z}$ and a ground truth condition $\boldsymbol{a}_g$, where each value in $\boldsymbol{a}_g$ corresponds with one specific attribute, e.g. stripes. While a discriminator $D(\boldsymbol{x}, \boldsymbol{a}_g)$ takes a pair of input features $\boldsymbol{x}$ and a ground truth condition $\boldsymbol{a}_g$ as input and a real value as output. Thus, the training objective of each local client model is defined as:

$$\mathcal{L}_{WGAN} = E[D(\boldsymbol{x}, \boldsymbol{a}_g)] - E[D(\tilde{\boldsymbol{x}}, \boldsymbol{a}_g)] - \lambda E[\left(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}}, \boldsymbol{a}_g)\|_2 - 1\right)^2] \quad (1)$$

$$\mathcal{L}_{CLS} = -E_{\tilde{\boldsymbol{x}} \sim p_{\tilde{\boldsymbol{x}}}}[\log P(y|\tilde{\boldsymbol{x}}; \theta)], \quad (2)$$

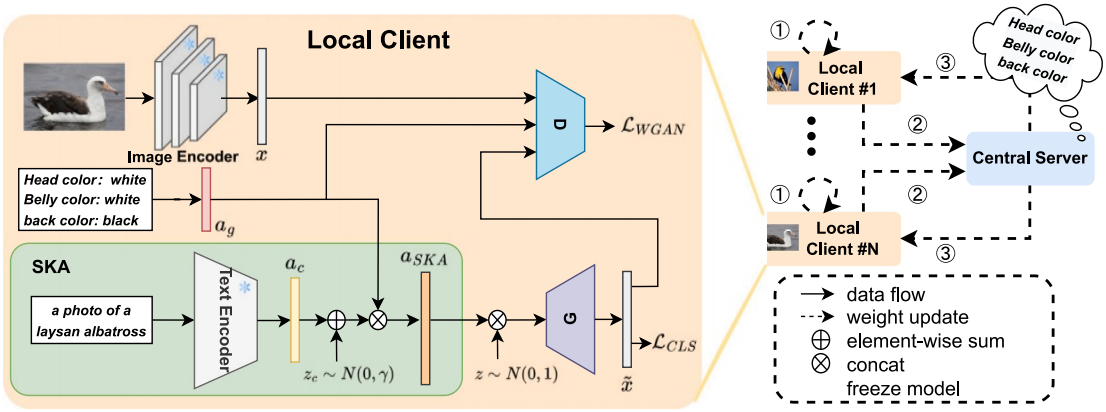


Fig. 2. An overview of federated zero-shot learning with mid-level semantic knowledge transfer. (1) Local model training process. (2) Local clients upload generator parameters to the server and the server constructs a global generator model by aggregating local generator parameters. (3) Local models are reinitialized with the central server model. The Semantic Knowledge Augmentation (SKA) employs external knowledge to further improve the model's discriminative ability.

$$\min_G \max_D \mathcal{L}_{WGAN} + \beta \mathcal{L}_{CLS}, \tag{3}$$

where $\boldsymbol{x}$ is CNN feature, $\tilde{\boldsymbol{x}} = G(\boldsymbol{z}, \boldsymbol{a}_g), \hat{\boldsymbol{x}} = \alpha \boldsymbol{x} + (1-\alpha)\tilde{\boldsymbol{x}}$, with $\alpha \sim U(0,1)$ and λ is the penalty coefficient. β is a hyper-parameter weight on the classifier, $\mathcal{L}_{WGAN}$ is Wasserstein GAN loss [30]. The conditional probability is computed by a linear softmax classifier parameterized by θ, which is pretrained on the real features of seen classes. $\mathcal{L}_{CLS}$ refers to the cross-entropy loss to classify generated CNN feature $\tilde{\boldsymbol{x}}$ to corresponding class label y.

After optimizing each local client model for E local epochs, the local model parameters, which include both generator parameters denoted as $\boldsymbol{w}_{G_i}$ and discriminator parameters $\boldsymbol{w}_{D_i}$ for client i, are transmitted to a central server for aggregation into a global model. To simplify, we denote the collection of generator and discriminator parameters in client i as $\boldsymbol{w}_i$, and in the server as $\boldsymbol{w}$. Following FedAvg [1], the aggregating process is formulated as:

$$\boldsymbol{w}_t = \frac{1}{N \cdot S} \sum_{i \in N_S} \boldsymbol{w}_{i,t}, \tag{4}$$

where N denotes the number of local clients and t denotes the tth global model iterative update round. S denotes the randomly selected clients fraction for each round ($S \in [0.0, 1.0]$) and N_S is the set of selected clients. Note that the central server only aggregates local model parameters without accessing local data so as to protect local data privacy. Then, each local model is reinitialized with the central model as:

$$\boldsymbol{w}_{i,t+1} = \boldsymbol{w}_t. \tag{5}$$

This is an iterative learning process (Eqs. (3)–(5)) until T global model update round. Since the attribute space is consistent among local clients, the learned global generator encodes mid-level semantic knowledge, i.e., given attribute vector $\boldsymbol{a}$ of a class, the generator can generate pseudo CNN features for this class regardless of whether it has been seen or not.

Testing Stage. In the testing stage, our objective is to classify images from unseen classes, where only the attributes of each unseen class are provided. To transfer from visual space to the label space, following [30], we utilize the trained generator from the global server to generate semantically rich CNN features conditioned on the attributes of unseen classes. These synthetic features are subsequently employed to train a discriminative softmax classifier, which is then directly tested on unseen classes.

Selective Module Aggregation. Although aggregating holistic model parameters following FedAvg [1] is simple, it is inefficient for FZSL because the generic mid-level semantic knowledge is mainly encoded in the generator while the discriminator may contain knowledge specific to classes in each client. Inspired by recent approaches [20,37] in federated learning, we improve the baseline by decoupling the discriminator from the central model aggregation process, i.e., only aggregating the generator in the central server. This not only reduces the cost for transmitting model parameters but also facilitates to learn more generalizable mid-level knowledge. Thus, the central aggregation in Eq. (4) and the local client reinitialization in Eq. (5) are reformulated as:

$$\boldsymbol{w}_{G,t} = \frac{1}{N \cdot S} \sum_{i \in N_S} \boldsymbol{w}_{G_i,t}, \tag{6}$$

$$\boldsymbol{w}_{G_i,t+1} = \boldsymbol{w}_{G,t}, \quad \boldsymbol{w}_{D_i,t+1} = \boldsymbol{w}_{D_i,t}, \tag{7}$$

where $\boldsymbol{w}_{G,t}$ and $\boldsymbol{w}_{D,t}$ denote model parameters for a generator and a discriminator, respectively.

Discussion. Note, we only aggregate the generator and decouple the discriminator in the selective module aggregation, which slightly differs from the process in [1,20,27]. Although selective module aggregation is not a significant contribution of this work, our experimental results show that the improved strategy outperforms the baseline. Thus, we use this useful strategy to improve a baseline model to facilitate the proposed mid-level semantic knowledge transfer for FZSL.

3.3. Semantic knowledge augmentation

Although the formulated baseline with selective module aggregation is able to transfer mid-level generic knowledge in a decentralized learning manner, it still suffers from sparse attribute and ambiguous attribute separability for limited data diversity in each client. To resolve this problem, we propose to explore a foundation model (language model (LM) RoBERTa [22] or text encoder of vision-language model (VLM) CLIP [23] in this work) to explore semantic knowledge augmentation (SKA) to enrich the mid-level semantic space in FZSL. Note we utilize off-the-shelf frozen text encoders directly which is disentangled from client–server aggregation process for efficient communication. Since a foundation model contains word embedding knowledge that can supply information regarding hierarchical relationships among classes, it can help FZSL to learn richer external knowledge with the sharable common attribute space. In this work, we introduce class-level semantic knowledge augmentation, which greatly facilitates the generated feature diversification in both training and testing stages. Empirically, we observe that directly concatenating a noise-enhanced text embedding and an attribute vector is an effective way, which do not require extra learnable parameters and can alleviate overfitting on seen classes.

In our semantic knowledge augmentation, as shown in Fig. 2, we simply combine a default prompt 'a photo of a' with class names and use this sentence as the input to the text encoder. We then further add the gaussian noise $\boldsymbol{z}_c \sim N(0,\gamma)$ to the output text embedding $\boldsymbol{a}_c$ so as to enrich the semantic space and to better align with the instance-wise diversified visual space, where each class-level semantic can always correspond to different samples with various poses and appearances in visual space. The semantic augmented attribute is the concatenation between noise-enhanced text embedding and ground truth manual annotation attribute labels $\boldsymbol{a}_g$. This semantic augmentation process is formulated as:

$$\boldsymbol{a}_{SKA} = [\boldsymbol{a}_c \oplus \boldsymbol{z}_c, \boldsymbol{a}_g], \tag{8}$$

where $\oplus$ is the element-wise summation. The local model training objective in Eqs. (1)–(3) is updated with

$$\tilde{\boldsymbol{x}} = G\left(\boldsymbol{z}, \boldsymbol{a}_{SKA}\right) \tag{9}$$

During FZSL model training, the text embedding of seen class name is utilized as external knowledge to construct semantic knowledge augmented attribute $\boldsymbol{a}_{SKA}$ and further generate image features in each local client. The discriminator condition keeps $\boldsymbol{a}_g$ to distinguish between the real distribution and the pseudo distribution. In the testing stage, instead of generating pseudo image features based on the same attribute $\boldsymbol{a}_g$ for each class as in conventional ZSL [30,38,39], the SKA module supplies diversified attribute $\boldsymbol{a}_{SKA}$ for each class. The gaussian noise $\boldsymbol{z}_c$ in $\boldsymbol{a}_{SKA}$ can help explore the rich information in foundation model text encoder so to enrich the attribute space. Overall, our semantic knowledge augmentation can increase inter-class separability and supply diversified attribute space by only using text information of the class name.

Algorithm 1 Federated Zero-Shot Learning with Mid-Level Semantic Knowledge Transfer.

Training stage:
initialize $\{w_{G_i,t=0}, w_{D_i,t=0}\}_{i=1}^{N}$
Server executes
for each round $t = 1$ to T **do**
  $N_S \leftarrow$ randomly select S fraction clients
  for each client $i \in N_S$ **do**
    $w_{G_i,t}, w_{D_i,t} \leftarrow$ **ClientUpdate**$(w_{G_i,t}, w_{D_i,t})$
  end for
  $w_{G,t} \leftarrow$ Eq. (6) ▷ Server aggregates generator
  $\{w_{G_i,t+1}, w_{D_i,t+1}\}_{i\in N_S} \leftarrow$ Eq. (7) ▷ Client models reinitialize
end for

ClientUpdate(w_G, w_D):
for local epoch $e = 0$ to $E-1$ **do**
  for mini-batch $b \subset S_k$ **do**
    $\boldsymbol{a}_{SKA} \leftarrow$ Eq. (8) ▷ Seen classes augmentation
    $w_G, w_D \leftarrow$ Eq. (1)- (3) with Eq. (9) ▷ Local training update G and D
  end for
end for
return w_G, w_D

Testing stage:
$\boldsymbol{a'}_{SKA} \leftarrow$ Eq. (8) ▷ Unseen classes augmentation
$f \leftarrow G(\boldsymbol{z}, \boldsymbol{a'}_{SKA}; w_{G,T})$ generate pseudo features for unseen classes
Softmax classifier F trained on pseudo features f
F is tested on unseen classes

Discussion. In this work, we consider a text encoder of a foundation model as a generic off-the-shelf model trained with a vast quantity of data. When using a text encoder of a vision-language foundation model (e.g., CLIP [23]), the text encoder has been trained with both texts and images so the model is exposed to plausibly similar images in training with the potential for zero-shot learning. However, since we only use the text encoder (not the whole vision-language model as PromptFL [35]) and the foundation model is defined as a generic off-the-shelf model, the text encoder can be used in our approach to explore semantic knowledge augmentation to enrich mid-level semantic space in FZSL. Besides, our approach can also be configured with a language model (e.g., RoBERTa [22]), which has not been trained with any images, to explore semantic knowledge augmentation for FZSL. Overall, with the improved baseline and semantic knowledge augmentation, our approach is capable of learning mid-level knowledge to facilitate FZSL. Since training data are non-shared and mid-level knowledge only contains semantic properties insensitive to objects of interests, our approach inherently protects privacy.

Table 1
Comparison with the related methods on AWA2, AWA1, aPY, CUB and SUN (Top-1 accuracy). SKA(LM) and SKA(VLM) denote semantic knowledge augmentation with text encoder from language model and vision-language model respectively. **Bold** and underline are the best and the second best results. Note that centralized methods are used for reference only because they are not direct counterparts.

	Method	AWA2	AWA1	aPY	CUB	SUN
Centralized	VAEGAN [38]	61.4	55.9	18.6	44.8	56.9
	CLSWGAN [30]	67.4	66.6	37.7	56.8	60.3
	FREE [39]	67.7	68.9	42.2	60.9	61.3
Decentralized	*Client number* $N = 4$					
	CLSWGAN+FedAvg [1]	61.6	58.5	33.8	53.8	59.5
	CLSWGAN+FedProx [9]	61.3	58.4	34.0	53.1	59.3
	CLSWGAN+MOON [8]	61.0	58.6	33.2	55.1	59.5
	VAEGAN [38]+FedAvg	60.2	51.6	18.7	43.2	55.5
	FREE [39]+FedAvg	60.9	59.8	25.9	54.5	56.4
	Ours (LM)	65.4	64.6	41.6	54.8	61.1
	Ours (VLM)	**69.0**	**70.6**	**47.1**	**59.4**	**66.5**
	Client number $N = 10$					
	FedZSL [28]	47.2	–	–	56.3	49.6
	Ours (VLM)	**68.7**	**67.1**	**41.6**	**59.7**	**67.1**

4. Experiments

Datasets. To evaluate the effectiveness of our approach, we conduct extensive experiments on five zero-shot benchmark datasets, including three coarse-grained datasets: (Animals with Attributes (AWA1), Animals with Attributes 2 (AWA2) and Attribute Pascal and Yahoo (aPY)); and two fine-grained datasets (Caltech-UCSD-Birds 200–2011 (CUB) and SUN Attribute (SUN)). AWA1 is a coarse-grained dataset with 30 475 images, 50 classes and 85 attributes, while AWA2 shares the same number of classes and attributes as AWA1 but with 37 322 images in total. The aPY dataset is a relatively small coarse-grained dataset with 15 339 images, 32 classes and 64 attributes. CUB contains 11 788 images from 200 different types of birds annotated with 312 attributes, while SUN contains 14 340 images from 717 scenes annotated with 102 attributes. We use the zero-shot splits proposed by [36] for AWA1, AWA2, aPY, CUB and SUN ensuring that none of training classes are present in ImageNet. All these five datasets are composed of seen classes set and unseen classes set. In decentralized learning experiments, we evenly split the seen classes set randomly to N clients. Note, both seen classes and unseen classes share the same attribute space in each dataset.

Evaluation Metrics. In FZSL, the goal is to learn a generalizable server model which can assign unseen class label $\mathcal{Y}_u$ to test images. Following commonly used zero-shot learning evaluation protocol [36], the accuracy of each unseen class is calculated independently before divided by the total unseen class number, i.e., calculating the average per-class top-1 accuracy of the unseen classes.

Implementation Details. Following [30,38,39], in our approach, we employed a frozen ResNet-101 pretrained on ImageNet as the CNN

feature extractor and constructed our baseline model with a generator and a discriminator for each client respectively following the representative zero-shot learning work [30]. Both generator and discriminator have a simple architecture consisting of only two MLP layers. Further, we employed a frozen pretrained text encoder (base RoBERTa [22] in SKA (LM) and text encoder of ViT-Base/16 CLIP [23] in SKA(VLM)) to supply class-name-based text embedding for each client. All clients share the same model structure while the server aggregates local model parameters to construct a global model. As for further improvement with semantic knowledge augmentation (SKA), the frozen text encoder is kept locally and not aggregated to the server. By default, we set the number of local clients $N = 4$, random client selection fraction $S = 1$ and noise augmentation γ to 0.1. Generated feature number M, penalty coefficient λ and classifier weight β follows [30]. We empirically set batch size to 64, maximum global iterations rounds $T = 100$, maximum local epochs $E = 1$. For each local client, we used Adam optimizer with a learning rate of $1e{-}3$ for CUB, $2e{-}4$ for SUN and $1e{-}5$ for the others. Our models were implemented with Python(3.6) and PyTorch(1.7) and trained on NVIDIA A100 GPUs.

4.1. Federated zero-shot learning analysis

There are no existing works discussing mid-level semantic knowledge transfer in federated learning, so besides our baseline model (CLSWGAN [30] with FedAvg [1]), we implement (1) a traditional ZSL method VAEGAN [38] and a recent ZSL method FREE [39] with FedAvg respectively, and (2) CLSWGAN with two other federated learning paradigms, namely FedProx [9] and MOON [8]. We simplify the notation by considering CLSWGAN+FedAvg interchangeable with *baseline* in the following section. All compared methods are inductive where only attribute information of unseen classes are used for training the classifier and unseen images are not used during training.

From Table 1, we can see that: (1) Compared with the centralized method (CLSWGAN), the decentralized baseline (CLSWGAN+FedAvg) yields compelling performance, showing the effectiveness of the proposed paradigm for learning globally generalized model whilst protecting local data privacy. This can be attributed to the learned generator which can effectively extract mid-level semantic knowledge in either decentralized settings with disjoint class label spaces or centralized settings. This observation brings inspiration to federated learning with non-IID data and proves that semantic mid-level information (attributes in particular) can be the consistent and generalizable knowledge extracted from heterogeneous clients so to boost all the participants. (2) Our approach significantly improves the baseline by 9% with SKA(VLM) and by 4% with SKA(LM) on average, which validates the effectiveness of our approach for FZSL. This can be attributed to the text encoder of foundation models which can supply rich semantic space and increase the diversity of generated pseudo samples to a large degree, which can further promote the generalizability of the model. (3) Comparing with other FL approaches (CLSWGAN+FedProx and CLSWGAN+MOON) and ZSL approaches (VAEGAN+FedAvg and FREE+FedAvg), our approach achieves significantly better performance on average, showing the adaptation of our method in learning mid-level semantic knowledge in federated learning. (4) We adapt the client number N to 10 and compare with FedZSL [28]. Our approach outperforms FedZSL by 14.13% on average, which demonstrates the effectiveness of our approach in federated zero-shot learning.

4.2. High-level vs. Mid-level knowledge transfer

Privacy Protection. To verify the effectiveness of privacy protection from transferring mid-level knowledge by attributes, we performed the gradient leakage attack [13] on federated learning with high-level knowledge transfer and our mid-level knowledge transfer. To simplify, we assume that the backbone of clients is LeNet, following [13]. The attacker can access both the backbone structure and the model gradient. As shown in Fig. 4, the source images used in high-level knowledge transfer are recovered pixelwise by the attacker, which cause privacy leakage. In contrast, in our mid-level knowledge transfer, only attribute vector can be recovered, which is a continuous embedding [36] and only leak less sensitive information. In extreme cases, even if the attribute-label mapping is further leaked to the attacker, the attacker can only know the corresponding categories (class-level) with millions of possibilities but not the exact object raw images (instance-level). This demonstrates high privacy security and robustness against privacy inference attack in our paradigm. To be explicit, our mid-level knowledge transfer use additional information, attribute feature, comparing to high-level knowledge transfer. This attribue feature can be considered as a quantized feature, which cannot be used to infer the original image as the gradient cannot be backpropagated to the original image space as a result of the discretization.

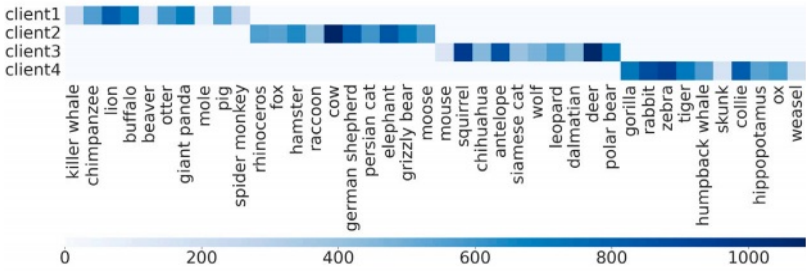
Model Generalization. As shown in Table 1, we evaluate various methods on an unseen label space to assess their model generalization abilities. Detailed quantitative analysis is discussed in Section 4.1. To supply a qualitatively analysis of model generalization from transferring mid-level knowledge by attributes, we visualize data partitions for FZSL on both label space and attribute space. In the perspective of class label partition as shown in Fig. 3(a), each client contains non-overlapping class label space which results to high degree of class label distribution skewness. Qu et al. [19] reported a notable performance decrease(76.25% on ResNet50) when comparing this extremely heterogeneous partition with IID data partition. However, in the perspective of attribute partition as shown in Fig. 3(b), it shows an overlapping attribute space and results to a lower degree of attribute label distribution skewness. Experiment results on Tables 1 and 2 prove the generalizability of transferring mid-level knowledge to unseen classes.

Table 2
Comparing local training and decentralized learning Top-1 accuracy in percentage on unseen classes. Baseline denotes CLSWGAN with FedAvg, while ours denotes Baseline+SMA+SKA (VLM).

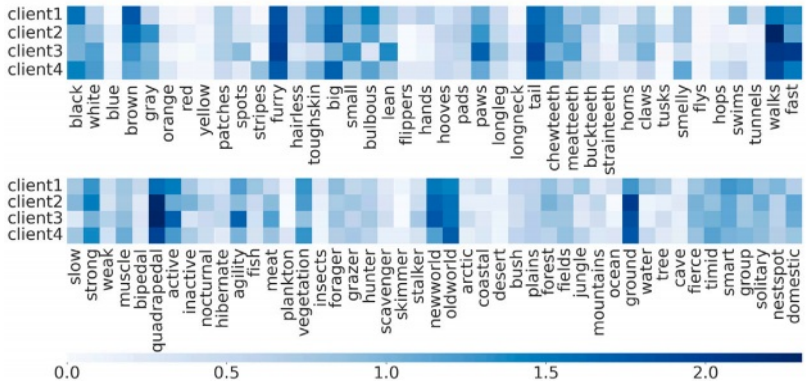
Setting	Methods	AWA2	AWA1	aPY	CUB	SUN
Local	Client 1	49.0	47.8	23.2	42.4	50.6
	Client 2	37.1	38.7	22.8	40.5	52.1
	Client 3	40.2	41.1	34.3	40.2	49.8
	Client 4	53.0	51.9	26.3	40.2	50.4
	Average	44.8	44.9	26.7	35.5	50.7
Decentralized	Baseline	61.6	58.5	33.8	53.8	59.5
	Ours (VLM)	69.0	70.6	47.1	59.4	66.5
Centralized		67.4	66.6	37.7	56.8	60.3
Centralized+SKA (VLM)		72.8	70.1	46.5	58.0	64.9

4.3. Local training vs. Decentralized learning

To verify the effectiveness of the formulated federated zero-shot learning paradigm, we separately train four individual local models with local client data and compare with decentralized learning models. For a clearer comparison, we further test the model in centralized learning with and without our proposed SKA. The performance are tested on the same unseen classes for all compared methods. As shown in Table 2, the decentralized baseline significantly outperforms all individual client models. This shows that the federated collaboration between the localized clients and the central server model facilitates to optimize a generalizable model in FZSL. Moreover, the shared mid-level knowledge among clients can be extracted from non-overlapping isolated local data with privacy protection and help boost the discriminative ability for unseen classes. Additionally, we evaluate SKA (VLM) with centralized method, which also shows a significant improvement and proves the effectiveness of proposed SKA in both centralized and decentralized scenario.



(a) Class label partition for each client on AWA2.



(b) Attribute partition for each client on AWA2.

Fig. 3. Class label vs. attribute partition for each client on AWA2. The depth of color represents the number of samples for each class in (a) and the sum of each continues attributes [36] for all classes in each client in (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

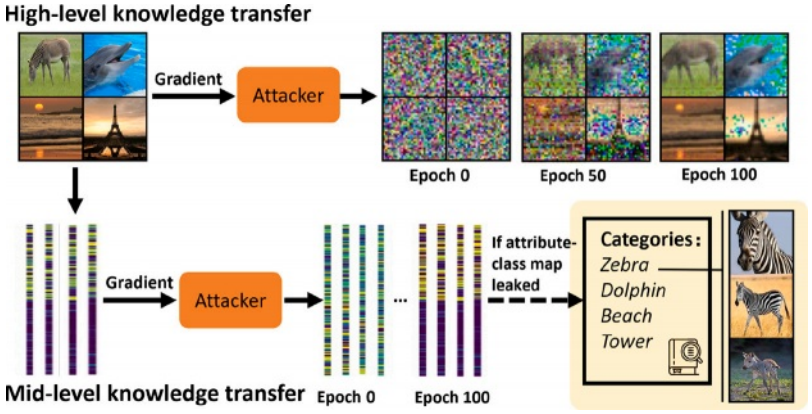


Fig. 4. The gradient leakage attack on high-level and mid-level knowledge transfer.

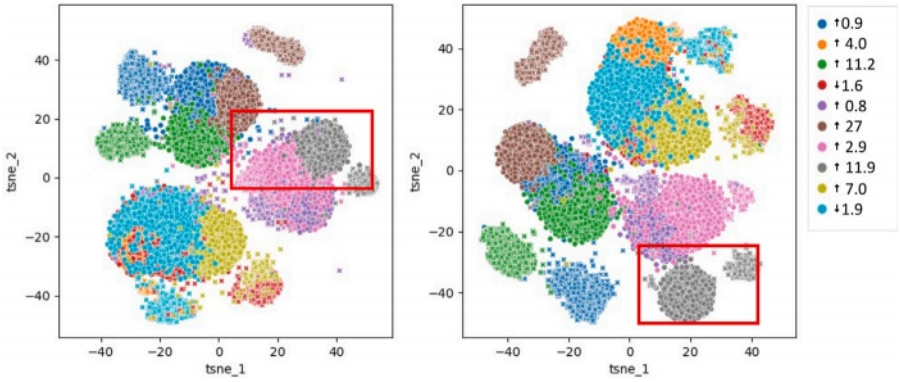


Fig. 5. tSNE of unseen classes on AWA2 for baseline+SMA (left) and baseline+SMA+SKA (VLM) (right). The same color implies the same class. Circle and cross means the generated distribution and real unseen distribution, respectively. The number in the caption means increase or decrease percentage for each class after implementing SKA (VLM). The classifier trained on generated pseudo distribution is tested on the unseen real distribution.

Table 3
Comparison with the related methods Top-1 accuracy on seen classes, unseen classes and harmonic mean H of the two accuracies.

	AWA2			CUB			SUN		
	Unseen	*Seen*	*H*	*Unseen*	*Seen*	*H*	*Unseen*	*Seen*	*H*
CLSWGAN+FedAvg	25.0	71.7	37.1	27.5	51.9	35.9	20.8	30.2	24.6
FREE+FedAvg	26.0	64.1	37.0	26.6	50.2	34.8	18.8	24.2	21.1
Ours (VLM)	**29.3**	**83.0**	**43.3**	**31.4**	**58.4**	**40.8**	**23.3**	**42.2**	**30.1**

Table 4
Baseline with proposed module variations. SMA: selective module aggregation. SKA: semantic augmentation with RoBERTa as language model (LM) and CLIP text encoder as vision-language model (VLM) respectively.

SMA	SKA	AWA2	AWA1	aPY	CUB	SUN
✗	✗	61.6	58.5	33.8	53.8	59.5
✓	✗	62.8	61.7	38.4	55.5	59.4
✗	✓(LM)	65.0	64.2	39.1	54.1	60.6
✓	✓(LM)	65.4	64.6	41.6	54.8	61.1
✓	✓(VLM)	69.0	70.6	47.1	59.4	66.5

Table 5
In comparison with Baseline+SMA, evaluation with the text embedding of two pretrained Language Models (LM) and two pretrained Vision-Language Models (VLM) are reported.

Text encoder		AWA2	AWA1	aPY	CUB	SUN
✗		62.8	61.7	38.4	55.5	59.4
LM	BERT	63.4	63.8	41.1	54.6	60.9
	RoBERTa	65.4	64.6	41.6	54.8	61.1
VLM	DeFILIP	74.1	75.5	49.4	58.2	64.2
	CLIP	69.0	70.6	47.1	59.4	66.5

Table 6
Evaluation with different CNN feature encoders.

Features	Methods	AWA2	CUB	SUN
GoogLeNet	CLSWGAN+FedAvg	55.4	45.4	51.1
	Ours (LM)	60.6	46.3	52.4
	Ours (VLM)	**65.2**	**50.9**	**55.8**
MobileNetV2	CLSWGAN+FedAvg	63.0	56.5	57.9
	Ours (LM)	66.3	57.1	58.1
	Ours (VLM)	**70.9**	**61.5**	**63.8**
ResNet101	CLSWGAN+FedAvg	61.6	53.8	59.5
	Ours (LM)	65.4	54.8	61.1
	Ours (VLM)	**69.0**	**59.4**	**66.5**

Table 7
Comparing baseline and ours (VLM) with both evenly and random unevenly split.

	Methods	AWA2	AWA1	aPY	CUB	SUN
Evenly	CLSWGAN+FedAvg	61.6	58.5	33.8	53.8	59.5
	Ours (VLM)	69.0	70.6	47.1	59.4	66.5
Unevenly	CLSWGAN+FedAvg	63.0	58.7	35.3	55.8	59.7
	Ours (VLM)	69.3	70.3	46.2	59.4	66.1

4.4. Federated generalized zero-shot learning

In this section, we evaluate the effectiveness of our approach in the federated generalized zero-shot learning (GZSL) scenario, where the model is evaluated to recognize samples from both seen and unseen classes simultaneously. We compare our approach with CLSWGAN [30] +FedAvg and FREE [39]+FedAvg, respectively. To adhere to local data protection requirements in federated learning, the server's generator is utilized to generate features for both seen and unseen classes. We compute the harmonic mean of the top-1 accuracy on seen classes and unseen classes following [30]. The experiments are conducted on three datasets: AWA2, CUB and SUN, which are widespread benchmarks for GZSL. As the results shown in Table 3, we observe that our method consistently outperforms both compared methods, which validates effectiveness of the proposed method in federated generalized zero shot learning scenario.

4.5. Ablation study

Component Effectiveness Evaluation. To evaluate the component of our approaches, We further analyze the results both quantitatively and qualitatively. Quantitatively, we report experimental results in Table 4 for the baseline with and without SMA, SKA (LM) and SKA (VLM) respectively. It can be observed that SMA can bring benefits (except baseline on SUN) with and without SKA, which can demonstrate the effectiveness of the SMA module to build an improved baseline for FZSL. Besides, SKA can bring significant improvement in all of the five datasets (except SKA(LM) in CUB), proving that the diversity supplied by SKA can help the global model generalizability. Qualitatively, the tSNE visualizations of AWA2 unseen classes for baseline+SMA before and after implementing the semantic knowledge augmentation with vision-language model are shown in Fig. 5. It can be seen that with SKA (VLM), the generated distribution has a larger inter-class distance as shown in the red box. This larger inter-class distance improves coarse-grained classification accuracy.

Variation of Semantic Knowledge Augmentation. In Table 5, we study more variants of text encoder in semantic knowledge augmentation. Here, BERT [34] is a bidirectional encoder similar to RoBERTa, while DeFILIP [40] is a variation of CLIP. As shown in Table 5, we can see that: (1) Both LM and VLM text encoder can bring benefits comparing with baseline, which shows the effectiveness and generalization capability of the proposed SKA structure. (2) Our approach with VLM achieves better results compared to ours with LM. The reason is mainly that the text encoders in VLM are pretrained with both texts and images so it is easier for these text encoders to achieve the alignment between visual and semantic distribution in FZSL. (3) DeFILIP, a fine-grained variation of CLIP, achieves the best result among different text encoders. Interestingly, we find that DeFILIP with attribute-level SKA (text input: 'a photo of a {attribute} {class name}.') can achieve 59.8% and 65.6% on CUB and SUN respectively (cf. 58.2% and 64.2% on CUB and SUN with class-level SKA), which implies that the fine-grained information from DeFILIP can be further explored by feature mining.

Variation of Encoder Architectures. In this work, following [30, 38,39], we mainly use ResNet101 as the CNN feature encoders. In Table 6, we further report the results of our approach with smaller CNN feature encoders, including GoogLeNet and MobileNetV2. From Table 6, we can observe that with CNN features extracted from different backbones, our approach can consistently improve the performance of the baseline. This shows that our method is not limited to ResNet-101 features. It is also applicable to other more compact and smaller network features.

Clients with Uneven Split. We further discuss the influence of data distribution with even or uneven class label split. In this work, we mainly conduct experiments with the evenly split client data distribution, where each client has equally number of non-overlapping class labels. Although even split of class labels for each client already represents an extremely non-IID data partition with high degree of label skewness [19], we further test on uneven split non-overlapping class labels. Specifically, the seen class labels are randomly split to 4 clients and ensure that each client has at least one-eighth of the total number of seen classes. For equal comparison, the baseline of

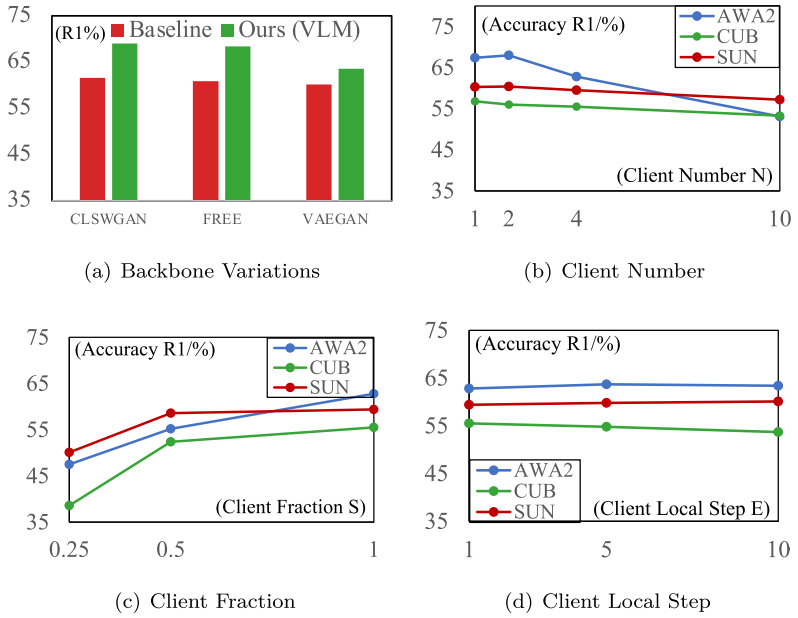


Fig. 6. Ablation study on (a) backbone variations, (b) client number, (c) client fraction, (d) local steps.

CLSWGAN+FedAvg and ours (VLM) are following exactly the same data split partition in both even and uneven setting. From Table 7, we can observe that: (1) The baseline can achieve comparable performance in both even and uneven split. This demonstrates that the extraction of mid-level knowledge is robust to the change in data distribution; (2) Our model performance is superior to CLSWGAN+FedAvg in both even and uneven splits, further validating the effectiveness and robustness of our method.

Backbone Variations. Fig. 6(a) compares FZSL with different backbone variations. We implemented backbones CLSWGAN [30], FREE [39], VAEGAN [38] with FedAvg [1] respectively, and further utilized our method with SKA(VLM) for comparison. We can see that our methods have a clear improvement on all of the three backbones. This validates the effectiveness and generalization capability of our method in FZSL.

4.6. Further analysis and discussion

In this section, we conduct the experiments on the improved baseline, namely CLSWGAN+FedAvg+SMA, to evaluate the effect of client number K, client fraction S and client local step E.

Client Number K. Fig. 6(b) compares central server aggregation with different numbers of local clients, where K = 1,2,4 and 10 represent seen classes of the dataset is randomly split to 1,2,4 and 10 clients on average respectively. We can see that the FZSL performance decreases when increasing the number of clients, which implies greater difficulty with larger number of clients with less data variety are used. But the performance degradation tendencies are stable on CUB and SUN, while that on AWA2 is more significant.

Client Fraction S. Fig. 6(c) compares FZSL with different client fractions. We can see that a smaller number of fraction is inferior to collaboration with a larger fraction of clients, which demonstrates that collaboration among multi-clients can further contribute to the generalization ability of the server model.

Client Local Step E. Fig. 6(d) compares FZSL with different client local steps E which influences the communication efficiency. Overall, the performance on different datasets shows relatively stable trends whilst on SUN, the performance decreases when E increases due to the accumulation of biases in local client.

5. Conclusion and future work

In this work, we proposed to exploit mid-level semantic knowledge transfer by attributes for federated learning, from which we introduced a new Federated Zero-Shot Learning paradigm. This mid-level knowledge transfer provides a generalized and privacy-protected solution for federated learning with data heterogeneity, particularly for clients with different label spaces but a sharable attribute space. We formulated a baseline model based on conventional zero-shot learning and federated learning. We further introduced selective module aggregation and foundation model based semantic knowledge augmentation to improve the generalization capability of the baseline model. Extensive experiments on five zero-shot learning datasets demonstrated our model's generalization ability as well as its effectiveness on privacy preserving.

As an early attempt for transferring mid-level knowledge in federated learning, our approach still has a number of limitations that need to be addressed in the future work. *First*, the mid-level attributes are predefined, leading to a higher requirement for data labeling. Therefore, developing automatic attribute label generation is one of the most effective methods. *Second*, the shareable mid-level attribute space is the same across clients and the server, which results in a limited number of use cases. How to deal with the situations where the attribute space is partially overlapping may need further investigation. We believe that with dedicated development in the above directions, the reliability of transferring mid-level knowledge in federated learning can be advanced significantly.

CRediT authorship contribution statement

Shitong Sun: Writing – review & editing, Writing, Visualization, Project administration, Methodology, Investigation, Formal analysis. **Chenyang Si:** Methodology. **Guile Wu:** Writing – review & editing. **Shaogang Gong:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the China Scholarship Council, the Alan Turing Institute Turing Fellowship, Veritone. We utilized Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.
[2] S. Sun, G. Wu, S. Gong, Decentralised person re-identification with selective knowledge aggregation, in: British Machine Vision Conference, 2021.
[3] J. Guo, H. Liu, S. Sun, T. Guo, M. Zhang, C. Si, FSAR: Federated skeleton-based action recognition with adaptive topology structure and knowledge distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10400–10410.
[4] H. Guan, P.-T. Yap, A. Bozoki, M. Liu, Federated learning for medical image analysis: A survey, Pattern Recognit. (2024) 110424.
[5] Z. Liu, F. Wu, Y. Wang, M. Yang, X. Pan, FedCL: Federated contrastive learning for multi-center medical image classification, Pattern Recognit. 143 (2023) 109739.
[6] B. Ma, Y. Feng, G. Chen, C. Li, Y. Xia, Federated adaptive reweighting for medical image classification, Pattern Recognit. 144 (2023) 109880.
[7] M. Ye, X. Fang, B. Du, P.C. Yuen, D. Tao, Heterogeneous federated learning: State-of-the-art and research challenges, ACM Comput. Surv. 56 (3) (2023) 1–44.
[8] Q. Li, B. He, D. Song, Model-contrastive federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10713–10722.
[9] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proc. Mach. Learn. Syst. 2 (2020) 429–450.
[10] T. Lin, L. Kong, S.U. Stich, M. Jaggi, Ensemble distillation for robust model fusion in federated learning, Adv. Neural Inf. Process. Syst. 33 (2020) 2351–2363.
[11] S. Wang, H. Tao, J. Li, X. Ji, Y. Gao, M. Gong, Towards fair and personalized federated recommendation, Pattern Recognit. 149 (2024) 110234.
[12] Z. Sun, X. Niu, E. Wei, Understanding generalization of federated learning via stability: Heterogeneity matters, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 676–684.
[13] L. Zhu, Z. Liu, S. Han, Deep leakage from gradients, Adv. Neural Inf. Process. Syst. 32 (2019).
[14] A. Majeed, Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data, J. King Saud Univ.-Comput. Inf. Sci. 31 (4) (2019) 426–435.
[15] A.L. Yuille, C. Liu, Deep nets: What have they ever done for vision? Int. J. Comput. Vis. 129 (3) (2021) 781–802.
[16] C.H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2013) 453–465.
[17] J. Chen, Y. Geng, Z. Chen, I. Horrocks, J.Z. Pan, H. Chen, Knowledge-aware zero-shot learning: Survey and perspective, in: International Joint Conference on Artificial Intelligence, 2021.
[18] Q. Yue, J. Cui, L. Bai, J. Liang, J. Liang, A zero-shot learning boosting framework via concept-constrained clustering, Pattern Recognit. 145 (2024) 109937.
[19] L. Qu, Y. Zhou, P.P. Liang, Y. Xia, F. Wang, E. Adeli, L. Fei-Fei, D. Rubin, Rethinking architecture design for tackling data heterogeneity in federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10061–10071.
[20] G. Wu, S. Gong, Collaborative optimization and aggregation for decentralized domain generalization and adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6484–6493.
[21] G. Wu, S. Gong, Decentralised learning from independent multi-domain labels for person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 2898–2906.
[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[23] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[24] X.-C. Li, D.-C. Zhan, 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, 2021, pp. 995–1005.
[25] J. Zhang, Z. Li, B. Li, J. Xu, S. Wu, S. Ding, C. Wu, Federated learning with label distribution skew via logits calibration, in: International Conference on Machine Learning, PMLR, 2022, pp. 26311–26329.
[26] W. Hao, M. El-Khamy, J. Lee, J. Zhang, K.J. Liang, C. Chen, L.C. Duke, Towards fair federated learning with zero-shot data augmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3310–3319.
[27] L. Zhang, D. Wu, X. Yuan, Fedzkt: Zero-shot knowledge transfer towards resource-constrained federated learning with heterogeneous on-device models, in: 2022 IEEE 42nd International Conference on Distributed Computing Systems, ICDCS, IEEE, 2022, pp. 928–938.
[28] Z. Chen, Y. Luo, S. Wang, J. Li, Z. Huang, Federated zero-shot learning for visual recognition, 2022, arXiv preprint arXiv:2209.01994.
[29] Y. Fu, T.M. Hospedales, T. Xiang, S. Gong, Transductive multi-view zero-shot learning, IEEE Trans. Pattern Anal. Mach. Intell. 37 (11) (2015) 2332–2345.
[30] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5542–5551.
[31] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, G. Peng, Conditional gaussian distribution learning for open set recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13480–13489.
[32] D.-W. Zhou, H.-J. Ye, D.-C. Zhan, Learning placeholders for open-set recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4401–4410.
[33] Z. Zhong, E. Fini, S. Roy, Z. Luo, E. Ricci, N. Sebe, Neighborhood contrastive learning for novel class discovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10867–10875.
[34] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.
[35] T. Guo, S. Guo, J. Wang, X. Tang, W. Xu, PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model, IEEE Trans. Mob. Comput. 23 (5) (2024) 5179–5194, http://dx.doi.org/10.1109/TMC.2023.3302410.
[36] Y. Xian, C.H. Lampert, B. Schiele, Z. Akata, Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2018) 2251–2265.
[37] T. Sheng, C. Shen, Y. Liu, Y. Ou, Z. Qu, Y. Liang, J. Wang, Modeling global distribution for federated learning with label distribution skew, Pattern Recognit. 143 (2023) 109724.
[38] Y. Xian, S. Sharma, B. Schiele, Z. Akata, F-vaegan-d2: A feature generating framework for any-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10275–10284.
[39] S. Chen, W. Wang, B. Xia, Q. Peng, X. You, F. Zheng, L. Shao, Free: Feature refinement for generalized zero-shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 122–131.
[40] Y. Cui, L. Zhao, F. Liang, Y. Li, J. Shao, Democratizing contrastive language-image pre-training: A CLIP benchmark of data, model, and supervision, 2022, arXiv:2203.05796.

Shitong Sun received the BEng degree in KU Leuven in 2017 and M.Sc. degree in KU Leuven in 2018. She is currently a Ph.D. student in Queen Mary University of London. Her current research interests include computer vision and pattern recognition.

Chenyang Si received the Ph.D. degree in University of Chinese Academy of Sciences (UCAS) in 2021. He is currently a research fellow at mmlab, Nanyang Technological University. His research interests lies at the intersection of deep learning and computer vision, including vision-based human perception.

Guile Wu received his Ph.D. degree in Queen Mary University of London.

Shaogang Gong is Professor of Visual Computation at Queen Mary University of London; elected a Fellow of the Royal Academy of Engineering (FREng), a Fellow of ELLIS, a Fellow of the Institution of Electrical Engineers, a Fellow of the British Computer Society, a member of the UK Computing Research Committee.