

Federated Zero-Shot Learning with Mid-Level Semantic Knowledge Transfer

Shitong Sun¹, Chenyang Si², Shaogang Gong¹, Guile Wu

¹Queen Mary University of London,

²Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences
shitong.sun@qmul.ac.uk, chenyang.si.mail@gmail.com, s.gong@qmul.ac.uk, guile.wu@outlook.com

Abstract

Conventional centralised 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 optimising a globally generalised central model (server). Existing federated learning paradigms mostly focus on transferring holistic high-level knowledge (such as class) across models, which are closely related to specific objects of interest so may suffer from inverse attack. 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 scalable. 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 generalised central model for deployment. To improve model discriminative ability, we propose to explore semantic knowledge augmentation from external knowledge for enriching the mid-level semantic space in FZSL. Extensive experiments on five zero-shot learning benchmark datasets validate the effectiveness of our approach for optimising a generalisable federated learning model with mid-level semantic knowledge transfer.

Introduction

Deep learning has gained great success in computer vision and natural language processing, but conventional deep learning paradigms mostly follow a centralised learning manner where data from different sources are collected to create a central database for model learning. With an increasing awareness of data privacy, decentralised deep learning (McMahan et al. 2017; Wu and Gong 2021b) is more desirable. To this end, federated learning (McMahan et al. 2017; Li et al. 2020b) has been recently introduced to optimise local models (clients) with non-shared local data while learning a global generalised central model (server) by transferring knowledge across the clients and the server. This enables to protect data privacy and reduce transmission cost 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 paradigms for computer vision applications, such as image classification (Chen and Chao 2021), person reidentifi-

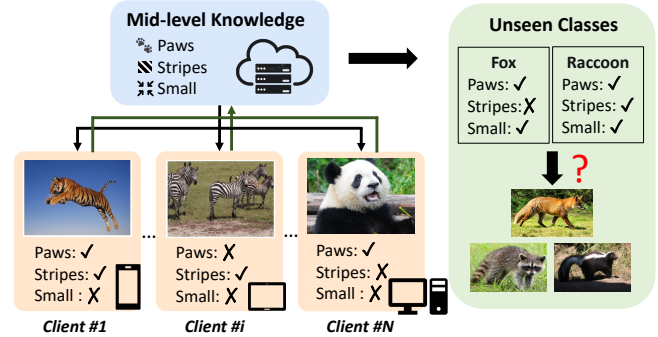


Figure 1: An overview of federated zero-shot learning with mid-level semantic knowledge transfer. Each local client optimises a local model with non-shared local data whilst a central server aggregates a global model by aggregating local model parameters. The server model will further be tested on unseen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

cation (Sun, Wu, and Gong 2021) and object detection (Liu et al. 2020).

However, existing federated learning paradigms (McMahan et al. 2017; Li et al. 2020b; Wu and Gong 2021b; Chen and Chao 2021) mostly focus on encoding holistic high-level knowledge into models for communication across the clients and the server. Since high-level knowledge is closely related to objects of interest, this may pose a threat to data privacy. In contrast, mid-level semantic knowledge (such as attribute) is usually generic containing semantically meaningful properties for visual recognition (Lampert, Nickisch, and Harmeling 2013), so it is not sensitive to objects of interests. Besides, since the number of attributes are finite in compositional learning (Yuille 2011) but the number of classes can be infinite, mid-level knowledge is also supposed to be more scalable. Therefore, learning mid-level semantic knowledge transfer for federated learning is important and is desirable for protecting privacy and improving model scalability.

On the other hand, zero-shot learning (ZSL) is a well-established paradigm for learning mid-level knowledge. It aims to learn mid-level semantic mapping between image features and text labels (typically attributes) using seen object categories and then transfer knowledge for recognising

unseen object categories with the help of the composition of shared attributes between seen and unseen categories. However, existing ZSL methods (Pourpanah et al. 2020; Chen et al. 2021a,b) mostly consider centralised learning scenarios which require to share training data from different label spaces to a central data collection.

In this work, we formulate a new Federated Zero-Shot Learning (FZSL) paradigm, which aims to learn mid-level semantic knowledge in federated learning for zero-shot learning in a decentralised learning manner. 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. Then, we optimise local models (clients) with non-shared local data and learn a central generalised model (server) by transferring knowledges (model parameters) between the clients and the server. With this paradigm, FZSL unifies federated learning and zero-shot learning for learning mid-level semantic knowledge in a decentralised learning manner with data privacy protection. It cumulatively optimises 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 (McMahan et al. 2017) or separating domain-specific classifiers like recent decentralized learning (Wu and Gong 2021a,b), we only aggregate generators across the clients and the server while discriminators are retained locally. This facilitates to learn more generalised knowledge and reduce the number of model parameters for communicating. Furthermore, to improve model discriminative ability, we employ a vision-language foundation model (e.g., CLIP (Radford et al. 2021)) to explore semantic knowledge augmentation to enrich the mid-level semantic space in FZSL. With the help of a pre-trained richer knowledge space, this semantic knowledge augmentation allows to learn a more generic knowledge to encode sample diversity as well as improve model scalability.

Our **contributions** are: We introduce a new Federated Zero-Shot Learning paradigm to transfer mid-level knowledge from independent non-overlapping class label spaces for federated learning. With the formulated baseline model, we propose to explore semantic knowledge augmentation from external knowledge to learn a richer mid-level semantic space in FZSL. We conduct extensive experiments on five zero-shot learning benchmark datasets and demonstrate that our approach is capable of learning a generalised federated learning model with mid-level semantic knowledge transfer.

Related Work

Federated Learning. Federated learning (McMahan et al. 2017; Li et al. 2020b,a) 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 many computer vision tasks, such as medical image segmentation (Liu et al. 2021), person reidentification (Wu and Gong 2021b), object detection (Liu et al. 2020), etc. Conventional federated learning approaches, e.g., FedAvg (McMa-

han et al. 2017), learn a sharable central model by aggregating holistic model parameters among different local models. To disentangle generic and specific knowledge, recent approaches (Wu and Gong 2021a; Zhang, Wu, and Yuan 2021; Wu et al. 2021; Sun, Wu, and Gong 2021) propose to optimise generic feature extractors or generators by decoupling discriminators or domain-specific classifiers, but are still learning holistic class-level knowledge. Different from existing works, we propose to learn mid-level semantic knowledge (i.e., attributes) for federated zero-shot learning. Although there have been several seemingly similar federated zero-shot learning studies (Gudur and Perepu 2021; Hao et al. 2021; Zhang, Wu, and Yuan 2021), none of these methods are aimed at bridging the gap between seen and unseen classes by learning mid-level semantic knowledge. ZSDG (Hao et al. 2021) generates existing categories by gathering statistics through the server. FedZKT (Zhang, Wu, and Yuan 2021) and (Gudur and Perepu 2021) are based on zero-shot knowledge distillation (Nayak et al. 2019) with the purpose of transferring knowledge between clients and server with no extracted prior information. Unlike them, our FZSL is learning from multiple independent *non-overlapping* class label spaces, while ZSDG (Hao et al. 2021) and (Gudur and Perepu 2021) are studying sharing knowledge with a sharing class space. Furthermore, our FZSL is generalisable and shows stable generalisability on *unseen* classes, while FedZKT and ZSDG are only tested on existing classes. More importantly, all of these methods are based on class-level knowledge while our FZSL learns to transfer mid-level semantic knowledge. Besides, we propose semantic knowledge augmentation from external knowledge to improve model discriminative ability for FZSL.

Zero Shot Learning. Zero shot learning (ZSL) aims to recognise 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 (Fu et al. 2015) and generative based methods (Xian et al. 2018b). 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 (Xian et al. 2018b). Existing ZSL methods are following a centralised 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.

Foundation Models. Foundation models refer to models trained with a vast quantity of data and can be further used for various downstream tasks, such as BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), CLIP (Radford et al. 2021), etc. These models are usually learned by self-learning using unlabelled data and are able to predict underlying properties such as attributes, so they are scalable and potentially more useful than models trained on a limited label

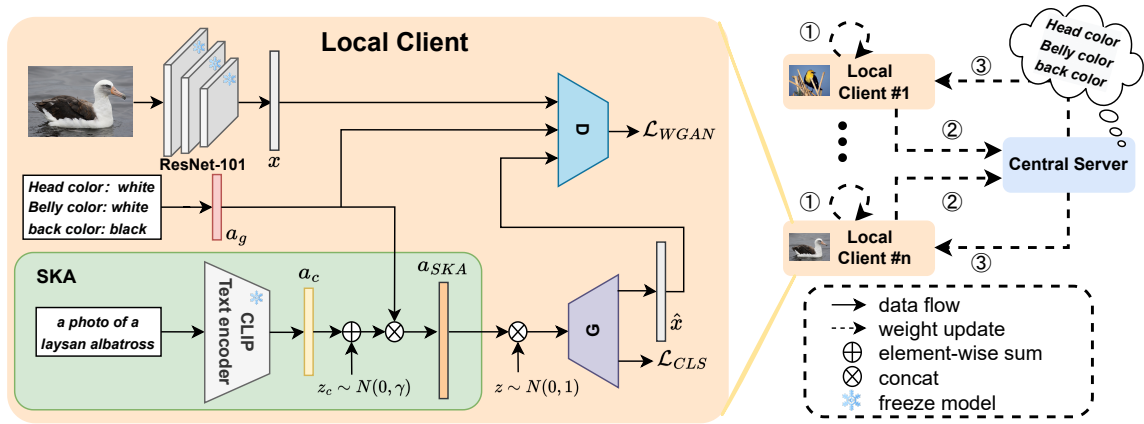


Figure 2: An overview of federated zero-shot learning with mid-level semantic knowledge transfer. (1) Local model training process. (2) Local clients upload model parameters to the server and server constructs a global model by aggregating local model parameters. (3) Local models are reinitialised with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

space (Bommasani et al. 2021). In this work, we employ a vision-language foundation model (e.g., CLIP (Radford et al. 2021)) to explore semantic knowledge augmentation enriching the mid-level semantic space in FZSL.

Methodology

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 while a central model is aggregated for deployment. Suppose there are N local clients, where the i -th client contains a training set $\mathcal{S}_i = \{\mathbf{x}, \mathbf{y}\}$, here $\mathbf{y} \in \mathcal{Y}_i$ includes N_i classes. Since each client contains non-overlapping class space, i.e., $\{\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset, \forall i, j\}$, $\mathcal{Y}_1 \cup \mathcal{Y}_2 \cup \dots \cup \mathcal{Y}_N = \mathcal{Y}_s$. Meanwhile, each class can be described by an attribute vector $\mathbf{a} = \{a_1, a_2 \dots a_m\}$ and these m attributes are consistent among classes in all clients, i.e. the mid-level attribute space is shared across clients. The goal of federated zero shot learning task is 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 and $\{\mathcal{Y}_i \cap \mathcal{Y}_u = \emptyset, \forall i, j\}$.

FZSL by Mid-Level Semantic Knowledge Transfer

A Baseline Model. 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 decentralised 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 f-CLSWGAN (Xian et al. 2018b) as the backbone (in practice, our approach is compatible to various ZSL backbones, such as VAEGAN (Xian et al. 2019) and FREE (Chen et al. 2021b)). As for federated learning, we use the commonly used FedAvg (McMahan et al. 2017). As shown in Fig. 2, the learning process of the baseline model consists

of three iterative steps, namely local model learning, central model aggregation and local model reinitialisation with central model.

In each local client, with the non-shared local data $\mathcal{S}_i = \{\mathbf{x}, \mathbf{y}\}$, the model learning process follows f-CLSWGAN (Xian et al. 2018b). A generator $G(\mathbf{z}, \mathbf{a}_g)$ learns to generate a CNN feature $\hat{\mathbf{x}}$ in the input feature space \mathcal{X} from random noise \mathbf{z} and a ground truth condition \mathbf{a}_g , where each value in \mathbf{a}_g corresponds with one specific attribute, e.g. stripes. While a discriminator $D(\mathbf{x}, \mathbf{a}_g)$ takes a pair of input features \mathbf{x} and a ground truth condition \mathbf{a}_g as input and a real value as output. Thus, the training objective of each local client model is defined as:

$$\min_G \max_D \mathcal{L}_{WGAN} + \beta \mathcal{L}_{CLS}, \quad (1)$$

where β is a hyper-parameter weight on the classifier.

After optimising each local client model for E local epochs, the local model parameters \mathbf{w}_i are transmitted to a central server to aggregate a global model. Following FedAvg (McMahan et al. 2017), the aggregating process is formulated as:

$$\mathbf{w}_t = \frac{1}{N \cdot S} \sum_{i \in N_S} \mathbf{w}_{i,t}, \quad (2)$$

where N denotes the number of local clients and t denotes the t -th 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 reinitialised with the central model as follows:

$$\mathbf{w}_{i,t+1} = \mathbf{w}_t. \quad (3)$$

This is an iterative learning process (Eqs.(1)-(3)) until T global model update round. Since the attribute space is consistent among local clients, the learned global generator encodes mid-level semantic knowledge. Finally, based on the

attributes from unseen classes, the learned generator from the global server is used to generate M pseudo image features for each unseen classes \mathcal{Y}_{un} . A softmax classifier is then trained under the supervision from pseudo features and tested for image classification on unseen classes.

Improved Baseline With Selective Module Aggregation.

Although aggregating holistic model parameters following FedAvg (McMahan et al. 2017) 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 (Wu and Gong 2021a; Zhang, Wu, and Yuan 2021) 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 generalisable mid-level knowledge. Thus, the central aggregation in Eq. (2) and the local client reinitialisation in Eq. (3) are reformulated as:

$$w_{G,t} = \frac{1}{N \cdot S} \sum_{i \in N_S} w_{G_i,t}, \quad (4)$$

$$w_{G_i,t+1} = w_{G,t}, \quad w_{D_i,t+1} = w_{D_i,t}, \quad (5)$$

where $w_{G,t}$ and $w_{D,t}$ denote model parameters for a generator and a discriminator, respectively.

Semantic Knowledge Augmentation for FZSL

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 vision-language foundation model (CLIP (Radford et al. 2021) in this work) to explore semantic knowledge augmentation (SKA) to enrich the mid-level semantic space in FZSL. Since a foundation model like CLIP 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 CLIP 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 a CLIP text encoder (Radford et al. 2021). We then further add the gaussian noise $z_c \sim N(0, \gamma)$ to the output text embedding 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 a_g . This semantic augmentation process can be formulated as follows:

$$a_{SKA} = [a_c \oplus z_c, a_g], \quad (6)$$

where \oplus is the element-wise summation. During FZSL model training, the CLIP text embedding of seen class name is utilised as external knowledge to construct semantic knowledge augmented attribute a_{SKA} and further generate image features in each local client. The discriminator condition keeps a_g to distinguish between the real distribution and the pseudo distribution.

Most importantly, in the testing stage, instead of generating pseudo image features based on the same attribute a_g for each class as in conventional ZSL (Xian et al. 2018b, 2019; Chen et al. 2021b), the SKA module supplies diversified attribute a_{SKA} for each class. The gaussian noise z_c in a_{SKA} can help explore the rich information in CLIP text encoder so to enrich the attribute space. Overall, our semantic knowledge augmentation can increase inter-class separability as well as supply diversified attribute space by only using the text information of the class name.

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) (Lampert, Nickisch, and Harmeling 2013), Animals with Attributes 2 (AWA2) (Xian et al. 2018a) and Attribute Pascal and Yahoo (aPY) (Farhadi et al. 2009)); and two fine-grained datasets (Caltech-UCSD-Birds 200-2011 (CUB) (Wah et al. 2011) and SUN Attribute(SUN) (Patterson and Hays 2012)). AWA1 is a coarse-grained dataset with 30475 images, 50 classes and 85 attributes, while AWA2 shares the same number of classes and attributes as AWA1 but with 37322 images in total. The aPY dataset is a relatively small coarse-grained dataset with 15339 images, 32 classes and 64 attributes. CUB contains 11788 images from 200 different types of birds annotated with 312 attributes, while SUN contains 14340 images from 717 scenes annotated with 102 attributes. We use the zero-shot splits proposed by (Xian et al. 2018a) for AWA1, AWA2, aPY, CUB and SUN ensuring that none of training classes are present in ImageNet (Russakovsky et al. 2015). All these five datasets are composed of seen classes set and unseen classes set. In decentralised learning experiments, we evenly split the seen classes set randomly to four 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 generalisable server model which can assign unseen class label \mathcal{Y}_u to test images. Following commonly used zero-shot learning evaluation protocol (Xian et al. 2018a), 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.

	Method	AWA2	AWA1	aPY	CUB	SUN
<i>Centralised</i>	CLSWGAN (Xian et al. 2018b)	67.4	66.6	37.7	56.8	60.3
	VAEGAN (Xian et al. 2019)	60.0	53.8	17.8	46.4	58.2
	FREE (Chen et al. 2021b)	67.7	68.9	42.2	60.9	61.3
<i>Decentralised</i>	CLSWGAN+FedProx (Li et al. 2020a)	61.3	58.4	34.0	53.1	59.3
	CLSWGAN+MOON (Li, He, and Song 2021)	61.0	58.6	33.2	55.1	59.5
	FL-VAEGAN	48.9	44.0	16.4	43.6	56.2
	FL-VAEGAN+SMA	<u>50.4</u>	<u>44.6</u>	25.9	<u>46.0</u>	<u>59.4</u>
	FL-VAEGAN+SMA+SKA	60.1	58.2	19.6	52.6	61.2
	FL-FREE	60.9	59.8	25.9	54.5	56.4
	FL-FREE+SMA	<u>61.4</u>	<u>61.1</u>	27.4	<u>55.4</u>	<u>57.0</u>
	FL-FREE+SMA+SKA	68.4	68.4	32.0	60.7	60.5
	FL-CLSWGAN	61.6	58.5	33.8	53.8	59.5
	FL-CLSWGAN+SMA	<u>62.8</u>	<u>61.7</u>	<u>38.4</u>	<u>55.5</u>	<u>59.4</u>
FL-CLSWGAN+SMA+SKA	69.0	70.6	47.1	59.4	66.5	

Table 1: Comparing our approach with other methods on AWA2, AWA1, aPY, CUB and SUN for federated zero-shot learning. Top-1 accuracy is reported on all experiments. SMA denotes selective module selection while SKA denotes semantic knowledge augmentation. **Bold** and underline represent the best and the second best performance in each baseline.

Implementation Details. In our approach, we employed a frozen ResNet-101 (He et al. 2016) pretrained on ImageNet (Russakovsky et al. 2015) as the feature extractor and constructed our baseline model with a generator and a discriminator for each client respectively following the representative generative zero-shot learning work (Xian et al. 2018b). Further, we employed a frozen pretrained CLIP (Radford et al. 2021) text encoder, a ViT-Base/16 transformer, 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. For the improved baseline with selective module aggregation (SMA), only the generator from local client are aggregated. As for further improved with semantic knowledge augmentation (SKA), both the generator and text-enhanced module are aggregated to the server. Each client contains local non-overlapping classes from the seen classes set and the aggregated server model is tested on the unseen classes set. By default, we set the number of local clients $N=4$ and randomly client select fraction $S=1$. Generated feature number M and classifier weight β follows the original ZSL work (Xian et al. 2018b). 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. Noise augmentation γ is set to 0.1 empirically. Our models were implemented with Python(3.6) and PyTorch(1.7), and trained on NVIDIA A100 GPUs.

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 (Xian et al. 2018b) with FedAvg (McMahan et al. 2017)) denoted as FL-CLSWGAN, we also implemented a traditional ZSL method VAEGAN (Xian et al. 2019) and a recent ZSL method FREE (Chen et al. 2021b) with FedAvg (McMahan et al.

2017) denoted as FL-VAEGAN and FL-FREE respectively for comparison. Further, the proposed SMA and SKA are implemented on three baselines respectively, where the generality and compatibility of SMA and SKA can be demonstrated. Note, when implementing SMA to FREE, feature refinement module will also be aggregated to the server which will be used during testing. 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 centralised baselines, the formulated decentralised baselines (FL-CLSWGAN, FL-VAEGAN, FL-FREE) yield compelling performance, which shows the effectiveness of the proposed paradigm for learning globally generalised model whilst protecting local data privacy; (2) With selective module selection (SMA), overall the performance of the baselines are improved (3.4% in FL-VAEGAN, 1% in FL-FREE and 2.1 % in FL-CLSWGAN on average), which verifies that learning a generic generator and decoupling the discriminator from central aggregation can facilitate mid-level semantic knowledge transfer in FZSL; (3) With semantic knowledge augmentation (SKA), our approach significantly improves the baselines by 8.5% in FL-VAEGAN, 6.5% in FL-FREE and 9.1% in FL-CLSWGAN on average, which validates the effectiveness and generality of SKA in FZSL; (4) Comparing with other federated learning approaches, such as FedProx (Li et al. 2020a) and MOON (Li, He, and Song 2021), our approaches achieve significantly better performance, showing the importance of learning mid-level semantic knowledge for FZSL. In the following context, the decentralised baseline donates CLSWGAN (Xian et al. 2018b) with FedAvg (McMahan et al. 2017) since it achieves overall the best performance on our experiments.

Local Training vs. Decentralised Learning

To verify the effectiveness of the formulated federated zero-shot learning paradigm, we separately train four individ-

Settings	Methods	AWA2	AWA1	aPY	CUB	SUN
Local Training	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
Decentralised	Baseline	61.6	58.5	33.8	53.8	59.5
	Baseline+SMA+SKA	69.0	70.6	47.1	59.4	66.5
Centralised	Baseline (Joint)	67.4	66.6	37.7	56.8	60.3

Table 2: Comparing local training (individual clients) and decentralised learning (baseline and baseline+SMA+SKA). Top-1 accuracy in percentage on unseen classes. Baseline donates CLSWGAN (Xian et al. 2018b) with FedAvg (McMahan et al. 2017)

GT	CLIP	AWA2	AWA1	aPY	CUB	SUN
✓	✗	62.8	61.7	38.4	55.5	59.4
✗	✓	70.1	72.4	48.2	42.2	54.4
✓	✓	69.0	70.6	47.1	59.4	66.5

Table 3: Baseline+SMA with different attribute variations. GT means dataset supplied annotated attributes. SKA means our proposed semantic augmentation with a CLIP text encoder.

(CL)SKA	ALSKA	AWA2	AWA1	aPY	CUB	SUN
✗	✗	62.8	61.7	38.4	55.5	59.4
✓	✗	69.0	70.6	47.1	59.4	66.5
✗	✓	62.8	64.4	44.8	54.4	61.6
✓	✓	69.3	70.7	46.2	59.0	65.6

Table 4: Baseline+SMA with different semantic augmentation variations. CLSKA means class-level semantic augmentation. ALSKA means attribute-level semantic augmentation.

ual local models (Xian et al. 2018b) with local client data and compare with decentralised learning models. Note that the performance are tested on the same unseen classes for all compared methods. As shown in Table 2, the decentralised baseline model significantly outperforms all individual client models and their average. This shows that the federated collaboration between the localised clients and the central server model facilitates to optimise a generalisable model in FZSL. Furthermore, baseline+SMA+SKA even surpasses the performance of the centralised joint-training baseline, which further verifies the effectiveness of our improved baseline for FZSL.

Effect of Semantic Knowledge Augmentation

As shown in Table 1, the performance of the baseline model can be significantly improved with semantic knowledge augmentation. To show the impact of semantic knowledge augmentation on FZSL, we further analyse the results both quantitatively and qualitatively. Quantitatively, we report experimental results in Table 3 for the baseline+SMA with and

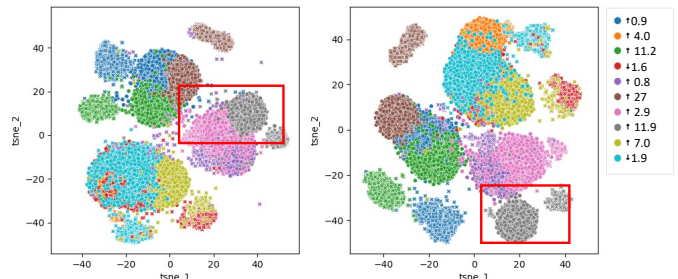


Figure 3: tSNE of unseen classes on AWA2 for baseline+SMA (left) and baseline+SMA+SKA (right). The same colour implies the same class. Circle and cross means the generated distribution and real unseen distribution, respectively. The number in the caption means increase or decrease percentage for each class after implementing SKA. The classifier trained on generated pseudo distribution is tested on the unseen real distribution.

without SKA. It can be observed from Table 3 that CLIP text embedding alone can supply discriminative information in three coarse datasets (AWA1, AWA2 and aPY) but lack discriminative ability in the other two fine-grained datasets. The combination of the ground truth annotation and CLIP text embedding, which is our SKA setting, works the best on average. Qualitatively, the tSNE visualisations of AWA2 unseen classes for baseline+SMA before and after implementing the semantic knowledge augmentation are shown in Fig. 3. It can be seen that with SKA, the generated distribution has a larger inter-class distance as shown in the red box. This larger inter-class distance significantly improves coarse-grained classification accuracy, which is consistent with the conclusion of FREE (Chen et al. 2021b).

Variation of Semantic Knowledge Augmentation

We do variations on the SKA in two directions: (1) In a more concrete attribute level and (2) text embedding from other text encoders.

Attribute-Level Semantic Knowledge Augmentation.

To further show whether an attribute text will bring more discriminative information to FZSL, we employ the attribute-

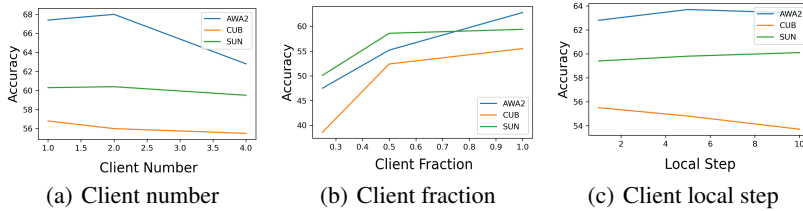


Figure 4: Ablation study on (a) client number, (b) client fraction, (c) local steps

Text Encoder		AWA2	AWA1	aPY	CUB	SUN
χ		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	54.7	61.0
VLP	DeFILIP	74.1	75.5	49.4	58.2	64.2
	CLIP	69.0	70.6	47.1	59.0	65.6

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

level semantic augmentation (ALSKA) and compare with the proposed class-level semantic augmentation ((CL)SKA). we reconstruct the input sentence of a CLIP text encoder with a superclass name and a random selected attribute from a target class. For example, for class ‘beach’, the input sentence can be constructed as ‘a photo of a swimming scene.’, where ‘scene’ is a superclass name and ‘swimming’ is a random selected positive attribute for class ‘beach’. Further, we combine ALSKA and (CL)SKA by constructing the input sentence of CLIP text encoder as ‘a photo of a {attribute} {class name}.’ where {attribute} is one of the activated attributes in {class name}. As shown in Table 4, we can see that: (1) Both class-level semantic augmentation (SKA) and the attribute-level semantic augmentation can supply discriminative information, which proves the effectiveness of our structure learning from text based external knowledge; (2) Comparing with (CL)SKA, the ALSKA is still limited in the CLIP text encoder. How to explore the fine-grained information from foundation model needs to be further explored and we leave this for the future work.

Semantic Knowledge Augmentation with Other Text Encoder. FZSL can gain benefit from a large scale pretrained text encoder. We naturally interested in whether other language models or visual language pretrained models can bring similar benefits. We therefore compare two large scale language models BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019); and the text encoder of a vision-language pretrained model DeFILIP (Cui et al. 2022). BERT and RoBERTa are bidirectional encoder trained on 16GB and 161GB text corpora respectively. DeFILIP is a variation of CLIP (Radford et al. 2021) which aims to explore fine-grained information in a more data efficient method. All of three methods will calculate the embedding of the whole input sentence, where we fed in the same sentence as our SKA.

As shown in Table 5, we can see that: (1) Both LM and VLP text encoder can bring benefits (except LM model on CUB) comparing with baseline, which can demonstrate the effectiveness and generality of the proposed SKA structure. (2) FZSL with VLP achieves better results compare to LM. The reason is mainly that these models are pretrained on image set and are prone to achieve the alignment between visual and semantic distribution. (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 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 with an appropriate mining method.

Further Analysis and Discussion

Client Number K . Fig. 4(a) compares central server aggregation with different numbers of local clients, where $K=1,2$ and 4 represent seen classes of the dataset is randomly split to 1,2 and 4 clients on average respectively. We can see that the FZSL performance decreases when implementing to increase number of clients, which implies greater difficulty with larger number of clients with less data variety.

Client Fraction S . Fig. 4(b) compares FZSL with different client fraction. We can see that a smaller number of fraction is inferior to collaboration with larger fraction of clients, which demonstrates that collaboration among multi-clients can further contribute to the generalisation ability of the server model.

Client Local Step E . Fig. 4(c) 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.

Conclusion

In this work, we introduced a new Federated Zero-Shot Learning paradigm to explore mid-level semantic knowledge transfer for federated learning. We formulate a baseline model based on conventional zero-shot learning and federated learning, and then further improve the baseline model with selective module aggregation and semantic knowledge augmentation. Extensive experiments on five zero-shot

learning benchmark datasets examine the effectiveness of our approach.

References

- Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Chen, H.-Y.; and Chao, W.-L. 2021. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*.
- Chen, J.; Geng, Y.; Chen, Z.; Horrocks, I.; Pan, J. Z.; and Chen, H. 2021a. Knowledge-aware zero-shot learning: Survey and perspective. *arXiv preprint arXiv:2103.00070*.
- Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; and Shao, L. 2021b. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 122–131.
- Cui, Y.; Zhao, L.; Liang, F.; Li, Y.; and Shao, J. 2022. Democratizing Contrastive Language-Image Pre-training: A CLIP Benchmark of Data, Model, and Supervision. *arXiv:2203.05796*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, 1778–1785. IEEE.
- Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2015. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(11): 2332–2345.
- Gudur, G. K.; and Perepu, S. K. 2021. Zero-Shot Federated Learning with New Classes for Audio Classification. *arXiv preprint arXiv:2106.10019*.
- Hao, W.; El-Khamy, M.; Lee, J.; Zhang, J.; Liang, K. J.; Chen, C.; and Duke, L. C. 2021. Towards fair federated learning with zero-shot data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3310–3319.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3): 453–465.
- Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10713–10722.
- Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020a. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2: 429–450.
- Li, X.; JIANG, M.; Zhang, X.; Kamp, M.; and Dou, Q. 2020b. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In *International Conference on Learning Representations*.
- Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P.-A. 2021. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1013–1023.
- Liu, Y.; Huang, A.; Luo, Y.; Huang, H.; Liu, Y.; Chen, Y.; Feng, L.; Chen, T.; Yu, H.; and Yang, Q. 2020. Fedvision: An online visual object detection platform powered by federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13172–13179.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.
- Nayak, G. K.; Mopuri, K. R.; Shaj, V.; Radhakrishnan, V. B.; and Chakraborty, A. 2019. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*, 4743–4751. PMLR.
- Patterson, G.; and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2751–2758. IEEE.
- Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C. P.; and Wang, X.-Z. 2020. A review of generalized zero-shot learning methods. *arXiv preprint arXiv:2011.08641*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.
- Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.
- Sun, S.; Wu, G.; and Gong, S. 2021. Decentralised Person Re-Identification with Selective Knowledge Aggregation. In *British Machine Vision Conference*.
- Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.
- Wu, G.; and Gong, S. 2021a. Collaborative optimization and aggregation for decentralized domain generalization and adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6484–6493.
- Wu, G.; and Gong, S. 2021b. Decentralised learning from independent multi-domain labels for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4, 2898–2906.

- Wu, Y.; Kang, Y.; Luo, J.; He, Y.; and Yang, Q. 2021. Fedcg: Leverage conditional gan for protecting privacy and maintaining competitive performance in federated learning. *arXiv preprint arXiv:2111.08211*.
- Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018a. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2251–2265.
- Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018b. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5542–5551.
- Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10275–10284.
- Yuille, A. L. 2011. Towards a theory of compositional learning and encoding of objects. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 1448–1455. IEEE.
- Zhang, L.; Wu, D.; and Yuan, X. 2021. FedZKT: Zero-Shot Knowledge Transfer towards Resource-Constrained Federated Learning with Heterogeneous On-Device Models. *arXiv preprint arXiv:2109.03775*.