南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

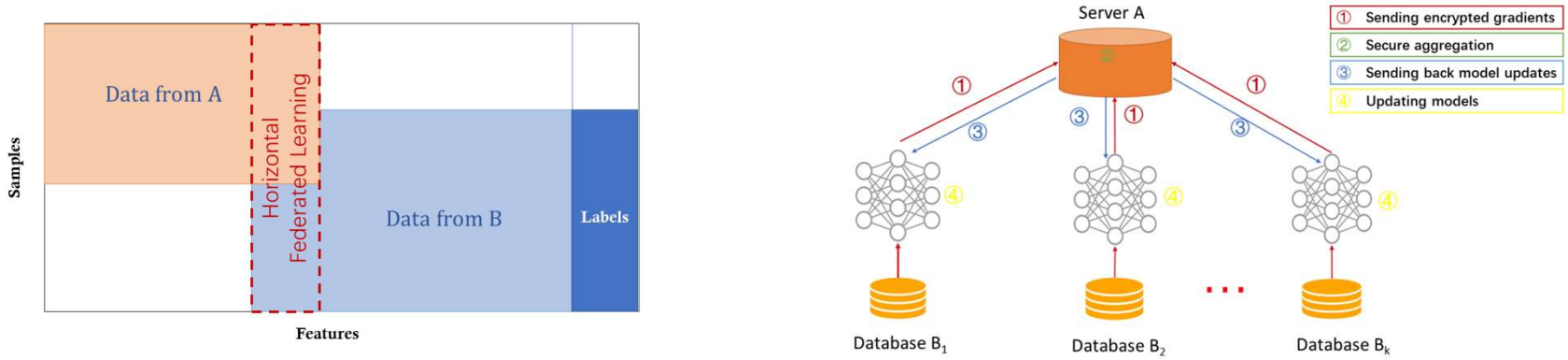
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

A Review of Federated Multi-Label Learning

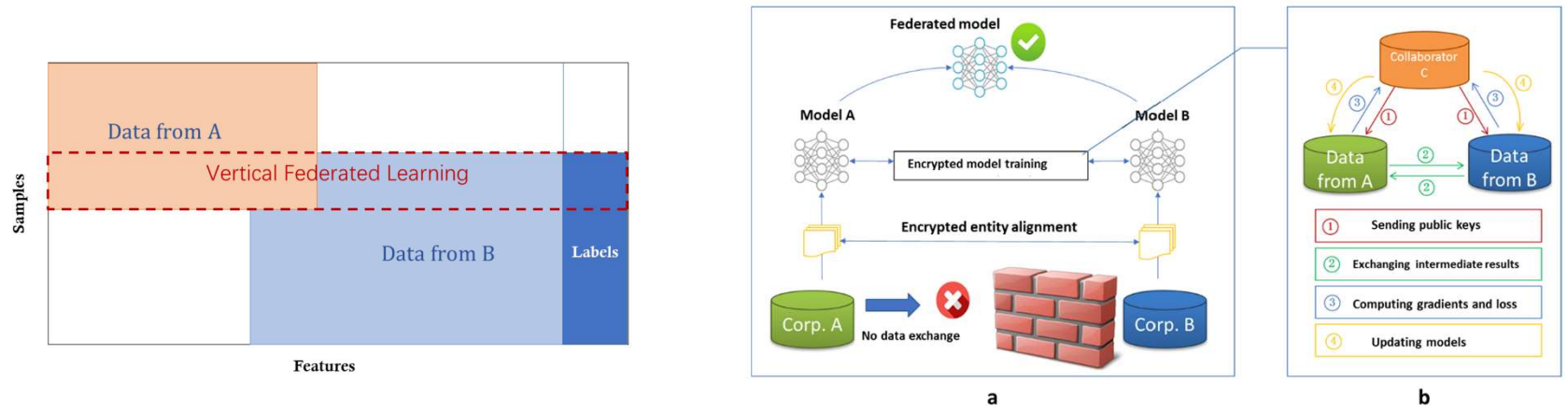
Reporter: Tong Jin

What is Federated Learning?

- Horizontal Federated Learning & Vertical Federated Learning



(a) Horizontal Federated Learning / Sample-Based Federated Learning
Datasets share the same feature space but different in samples.



(b) Vertical Federated Learning / Feature-Based Federated Learning
Datasets share the same sample ID but differ in feature space.

Eg. Two clients, Skin lesion dataset (100 samples (sample ID:1-100, two features: dermoscopic image, metadata))

Challenges in Federated Learning

- **Heterogeneity：**
  - ✓ **Data heterogeneity**
    - label distribution skew
    - quantity skew
    - domain shifts
  - ✓ Model heterogeneity
  - ✓ Resources heterogeneity
- Privacy protection
- Communication efficiency

All of these are designed for single-label classification tasks.

Federated Multi-label classification: Similar issues persist and can co-occur!

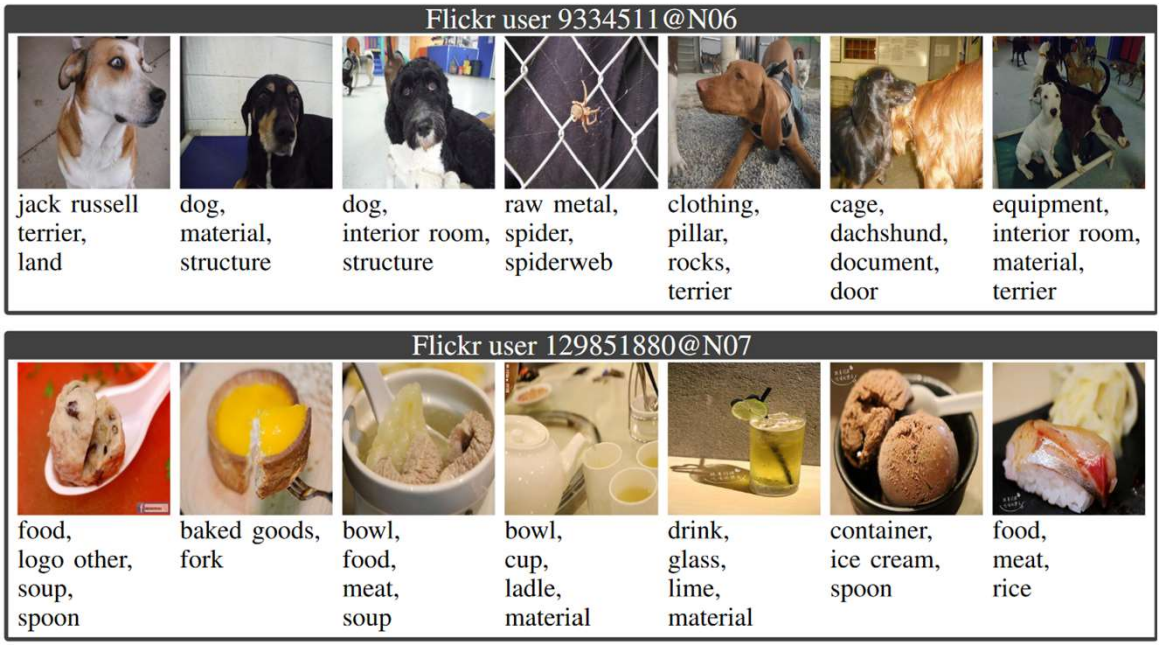
FLAIR : A Real-World Dataset Designed for Multi-Label FL

FLAIR has 429,078 images from 51,414 Flickr users. Two levels of difficulty: the easier task has 17 coarse-grained classes and the harder task has 1,628 fine-grained classes.

Quantity skew: Users have different number of images. The majority of users have only 1-10 images, but the most active users have hundreds of images.

Domain shifts: Users have different cameras, camera settings, which affect pixel generation.

Label distribution skew: Users take photos of objects that align with their interests, which vary across photographers.



Existing Work

- Federated Learning for Site Aware Chest Radiograph Screening (IEEE ISBI 2021)
- Federated Partially Supervised Learning with Limited Decentralized Medical Images (IEEE TMI 2023)
- FLAG: Fast Label-Adaptive Aggregation for Multi-Label Classification in Federated Learning (arxiv 2023)
- Federated Learning with Only Positive Labels by Exploring Label Correlations (TNNLS 2024)
- Scene-based Graph Convolutional Networks for Federated Multi-Label Classification (IJCNN 2024)
- **Language-Guided Transformer for Federated Multi-Label Classification (AAAI 2024)**

Federated Learning for Site Aware Chest Radiograph Screening

Problem: The large variations in disease prevalence and co-morbidity distributions (疾病患病率和共病分布) across the sites may hinder proper training.

Method:

Training separate CNN models at each site may lead to **over-fitting** due to the **small training sample size** at each site
→ **Aggregate CNN model weights** to learn robust site-independent features.

A global CNN model trained using FL is unable to capture the **local site-specific variations in the data distributions**
→ **Constructe separate GNN** to leverage the local site-specific disease prevalence and co-morbidity statistics for disease classification.

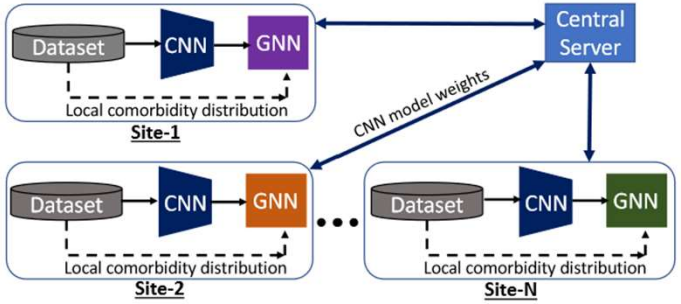


Fig. 1. A sequence of CNN-GNN model is trained across N sites. A global CNN model weight is shared across all sites while separate GNN model weights are learned at each site to capture the local co-morbidity dependencies among the disease classes. The central server periodically receives the local updates of the CNN weights from each site, aggregates them and sends back the global CNN weights to each site using Federated Learning.

Disadvantage: Label relationships do not have global interactions, and only using local limited data may lead to the problem of label relationship overfitting.

FLAG: Fast Label-Adaptive Aggregation for Multi-Label Classification in Federated Learning

Method:

Clustering-based Multi-label Data Allocation: According to the label distribution of the samples, a number of clusters are formed by *K*-modes clustering.

Assign samples to different clusters by minimizing its dissimilarity measure to the center. For two sample X and Y with m categorical features, the dissimilarity measure is

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

$$\delta(x_j, y_j) = \{1, x_j = y_j; 0, x_j \neq y_j\}$$

Cluster center : (1, 0, 1, 1)
Sample label distribution: (0, 1, 1, 1)

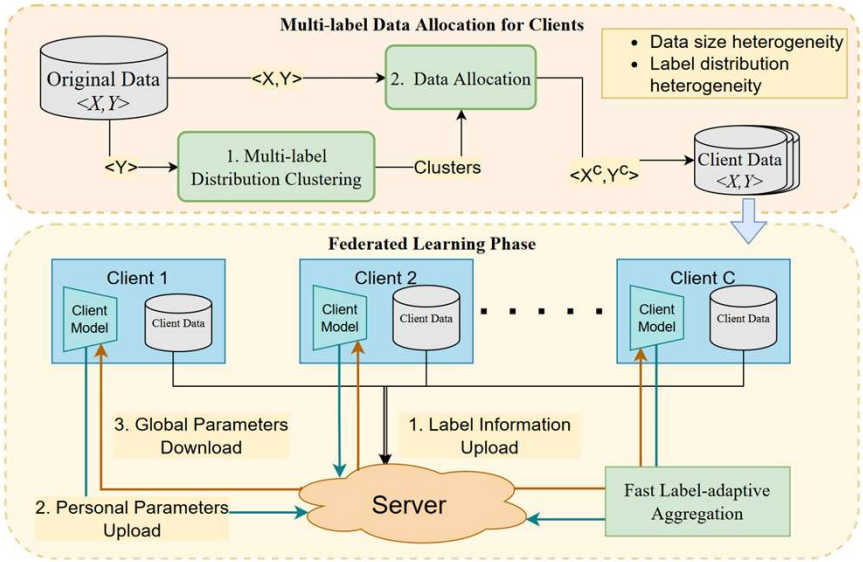


Figure 1: The multi-label federated learning framework

Fast label-adaptive Aggregation: Compute the aggregation weight considering both label distribution and label occurrence.

$$label\_weight = \{\sum_{l=1}^{N_l} (\sum_{i=1}^{N_i^c} y_i^c)^{\alpha}, 1 \leq c \leq N_c\}$$

α controls the importance between label occurrence and distribution. α = 0 means label weight only considers label distribution wideness.

Federated Partially Supervised Learning with Limited Decentralized Medical Images

Setting: Federated partially supervised learning (FPSL)

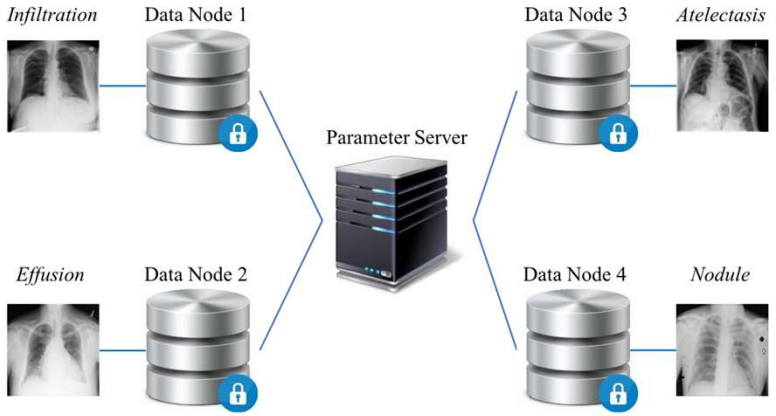


Fig. 1. Illustration of FPSL for a multi-label classification task on chest X-ray images. Here, each client (data node) is annotated for only one thoracic disease. We use this simple example to convey the main concept of the problem of interest (in practice, each client could be partially labeled for multiple classes). In this scenario, we only know whether each image in the first data node has *infiltration* but have no knowledge on the other three diseases. To ensure data government, only model weights and the metadata (*e.glet@tokeneonedot* statistics) of the local data can be communicated between each data node and the parameter server (see Sec. IV for a formal description). The goal of FPSL is to utilize the four partially labeled datasets stored in the different data nodes to train the model of interest in the parameter server.

Challenges of Federated Partially Supervised Learning

A Multi-Task Representation

Multi-label predictor: $g_\phi \circ f_\theta$

Given an input x, we have: $(g_\phi \circ f_\theta)(x) = g_\phi(f_\theta(x))$

ϕ can be further represented as a $C \times d$ matrix, where d is the dimension of the output feature vector. We can represent the classification network output as

$$(g_\phi \circ f_\theta)(x) = \phi \cdot f_\theta(x)$$

Given C classes of interest, we can decompose the weight matrix ϕ into C weight vectors

$$\phi = \begin{bmatrix} \phi^1 \\ \phi^2 \\ \vdots \\ \phi^C \end{bmatrix}$$

For a single class c, the probability score of the prediction is then

$$p_c(x) = \text{sigmoid}(\phi^c \cdot f_\theta(x)),$$

Challenges of Federated Partially Supervised Learning

Simple FedAVG

Global prediction model

$$\theta_0 = \sum_k w_k \theta_k,$$

$$\phi_0 = \sum_k w_k \phi_k,$$

$$w_k = \frac{n_k}{\sum_k n_k} \text{ and } \sum_k w_k = 1$$

look at a single class c

$$\begin{aligned} g_{\phi_0^c} \circ f_{\theta_0}(x) &= \phi_0^c \cdot f_{\theta_0}(x) \\ &= \sum_k w_k \phi_k^c \cdot \sum_k w_k f_{\theta_k}(x) \\ &= \sum_{i,j} w_i w_j \phi_i^c \cdot f_{\theta_j}(x). \end{aligned}$$

clients are split into two sets, which are $\mathcal{K}_L$ and $\mathcal{K}_U$.

$$\begin{aligned} g_{\phi_0^c} \circ f_{\theta_0}(x) = & \sum_{i_L \in \mathcal{K}_L, j_L \in \mathcal{K}_L} w_{i_L} w_{j_L} \phi_{i_L}^c \cdot f_{\theta_{j_L}}(x) \\ & + \sum_{i_L \in \mathcal{K}_L, j_U \in \mathcal{K}_U} w_{i_L} w_{j_U} \phi_{i_L}^c \cdot f_{\theta_{j_U}}(x) \\ & + \sum_{i_U \in \mathcal{K}_U, j_L \in \mathcal{K}_L} w_{i_U} w_{j_L} \phi_{i_U}^c \cdot f_{\theta_{j_L}}(x) \\ & + \sum_{i_U \in \mathcal{K}_U, j_U \in \mathcal{K}_U} w_{i_U} w_{j_U} \phi_{i_U}^c \cdot f_{\theta_{j_U}}(x). \end{aligned}$$

For unlabeled clients K_U (with respect to class c), no contributions should be made to the model aggregation as no label information are utilized in the local training. Thus, the third and the fourth terms will inevitably degrade the final performance.

Method

Task-Dependent Model Aggregation

$$\theta_0 = \sum_k \frac{n_k}{\sum_k n_k} \theta_k,$$
$$\phi_0^c = \sum_k \frac{n_k^c}{\sum_k n_k^c} \phi_k^c,$$

where n_k^c denotes the number of labeled examples in client k with respect to class c.

Task-Agnostic Decoupling Learning

Rephrase local training target as a bi-level optimization problem. Assume the local training data can be split into two subsets, one is denoted as the training set and the other one is denoted as the validation set.

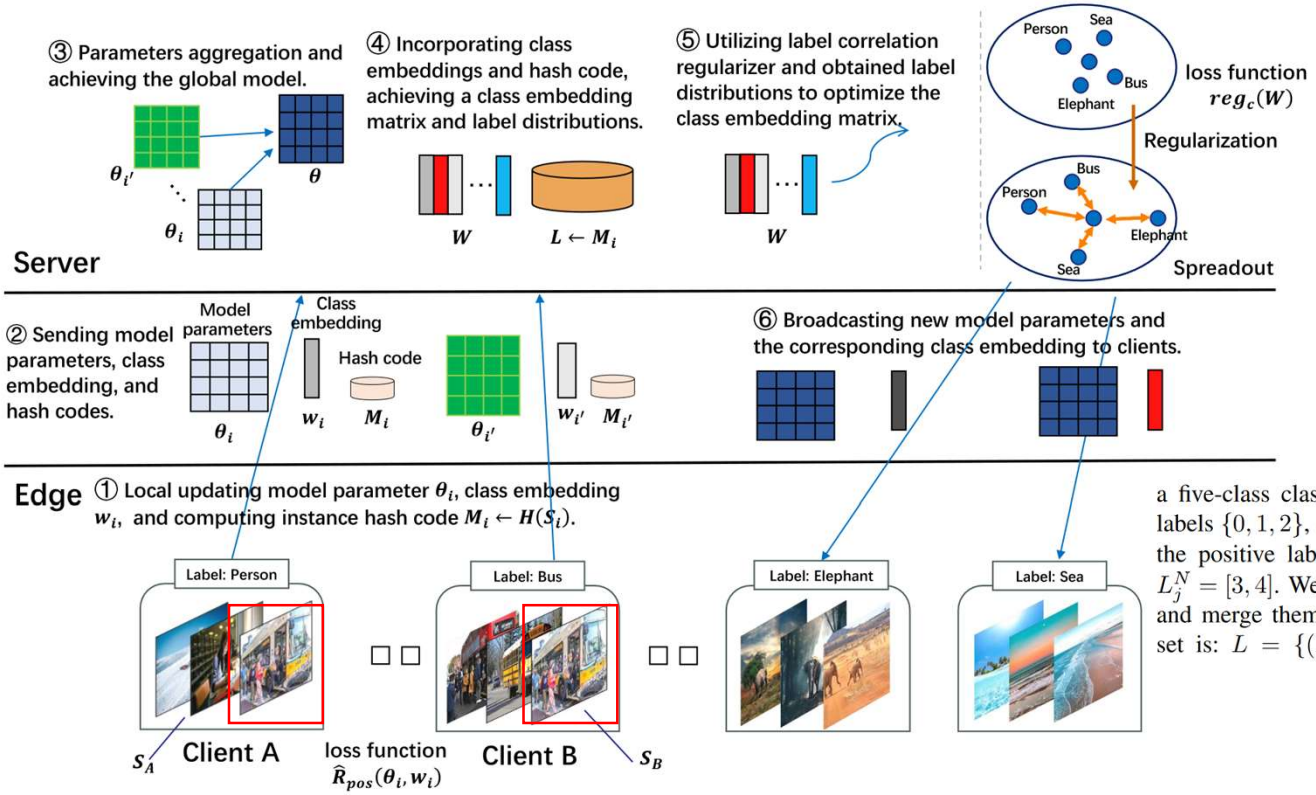
$$\min_{\theta} \mathcal{L}_{val}(g_{\phi^*|\theta} \circ f_\theta(x), y);$$
$$\text{s.t. } \phi^*|\theta = \arg\min_{\phi} \mathcal{L}_{train}(g_\phi \circ f_\theta(x), y),$$

where θ is the upper level variable and ϕ is the lower level variable. Intuitively, we misalign the sample spaces of f_θ and g_ϕ to mitigate local overfitting.

Due to partial supervision, the quality of f in a client could be **biased to a few classes**. While this may have a positive effect on the first term for class c, the other classes can be negatively influenced by the last three terms. However, with limited data, the overfitting could be severe.

Federated Learning with Only Positive Labels by Exploring Label Correlations

Setting: only positive data w.r.t. a single class label is provided for each client.



$$\mathrm{reg}_{\mathrm{fc}}(W) = \sum_{i\in[m]} \frac{|\mathcal{S}^i|}{|\mathcal{S}|}\left(\left(\hat{\mathcal{R}}_{\mathrm{pos}}(\mathcal{S}^i) + \hat{\mathcal{R}}_{\mathrm{neg}}(\mathcal{S}^i)\right)\right.$$

$$= \frac{1}{n_{all}} \sum_{i\in[m]} \sum_{j\in[n_i]} \left(\ell_{cl}^{\mathrm{pos}}(L_j^P, L_j^N, W) + \ell_{cl}^{\mathrm{neg}}(L_j^P, L_j^N, W)\right)$$

$$= \frac{1}{n_{all}} \sum_{j\in[n_{\mathrm{all}}]} \left(\alpha \cdot \sum_{\substack{y_j\in L_j^P, \\ y_j'\in L_j^P \,\&\, y_j'\neq y_j}} \left(d(\mathbf{w}_{y_j}, \mathbf{w}_{y_j'})\right)^2 \right.$$

$$\left. + \beta \cdot \sum_{\substack{y_j\in L_j^P, \\ y_j'\in L_j^N}} \left(\max\left\{0, \nu - d(\mathbf{w}_{y_j}, \mathbf{w}_{y_j'})\right\}\right)^2 \right).$$

a five-class classification problem, there is an instance with labels $\{0, 1, 2\}$, and then its label set is $L_j = [L_j^P, L_j^N]$, where the positive label set $L_j^P = [0, 1, 2]$ and negative label set $L_j^N = [3, 4]$. We collect the label sets of all clients' instances and merge them on the server, and the resulting whole label set is: $L = \{(L_1^P, L_1^N), (L_2^P, L_2^N), ..., (L_{n_{\mathrm{all}}}^P, L_{n_{\mathrm{all}}}^N)\}$

Fig. 1. **Overview of the proposed federated averaging by exploring label correlations (FedALC) method**. **(1)** Our FedALC computes gradients for parameter updating and hash code for each instance locally. The instance hash code is utilized for calculating label correlations and **only need transmission once**; **(2)** The client sends the locally updated model parameters, class embedding, and hash codes to the server; On the server, **(3)** the global model is obtained via parameter aggregation; **(4)** the different class embeddings are merged as a matrix, and label distribution is obtained by comparing the hash codes; **(5)** The server then utilizes our designed correlation regularizer based on the label distribution to optimize the class embedding matrix; **(6)** and eventually transmits global model parameters and corresponding class embeddings to different clients.

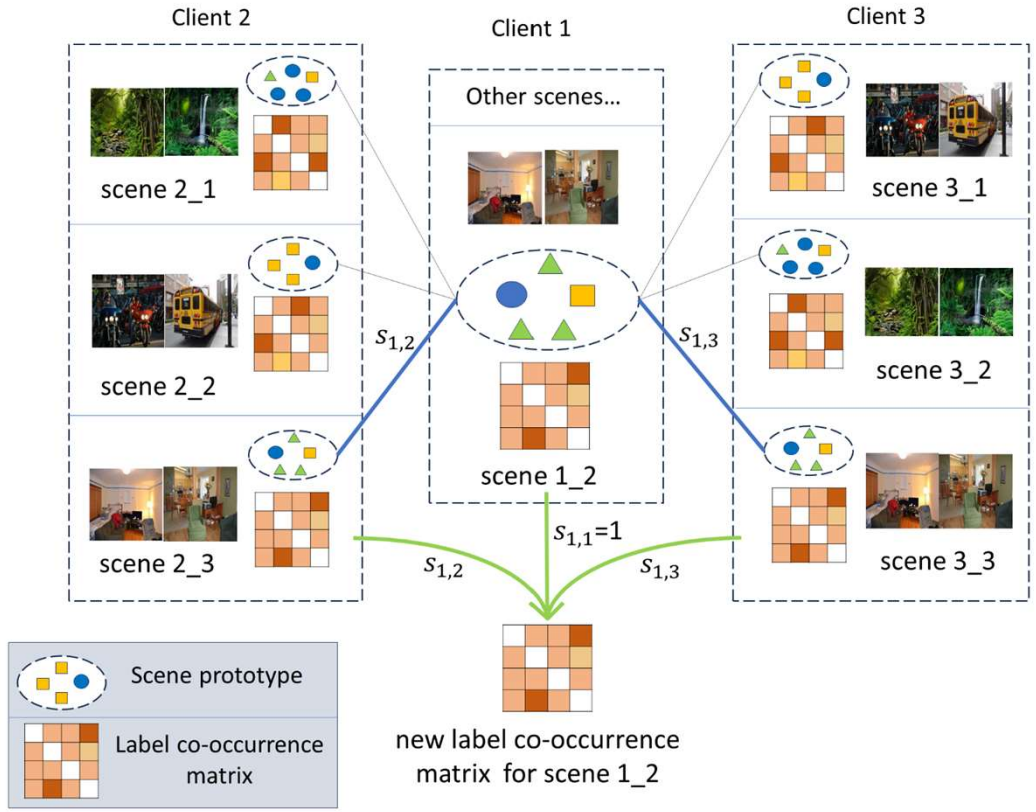


Fig. 2: The illustration of the proposed method for aggregating the label co-occurrence matrices. Firstly, we calculate the similarities between the scene prototypes of different clients. Then the most similar scenes from other clients are identified, and their label co-occurrence matrices are aggregated with the aggregation weight being the similarities of the scene prototypes. Specifically, for the second scene of client 1, it is found that the third scene of client 2 and the third scene of client 3 are the most similar scenes. Therefore, the co-occurrence matrices of these scenes are aggregated. Since the similarities between clients' scenes are not necessarily symmetric, the above process is repeated for each scene of every client.

Language-Guided Transformer for Federated Multi-Label Classification

FedLGT is the first to tackle the problem of label discrepancy across different clients for multi-label FL.

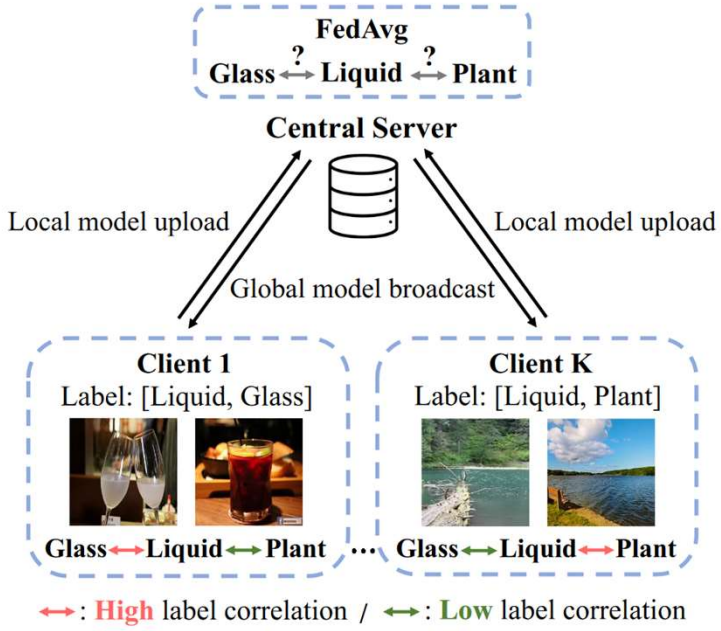


Figure 1: Challenges in multi-label federated learning. Since diverse label correlations are observed across clients, aggregating local models might not be sufficiently generalizable.

Why multi-label learning cannot be directly applied: Traditional centralized multi-label learning methods can model global view of label relationships, which is infeasible under FL scenarios.

Language-Guided Transformer for Federated Multi-Label Classification

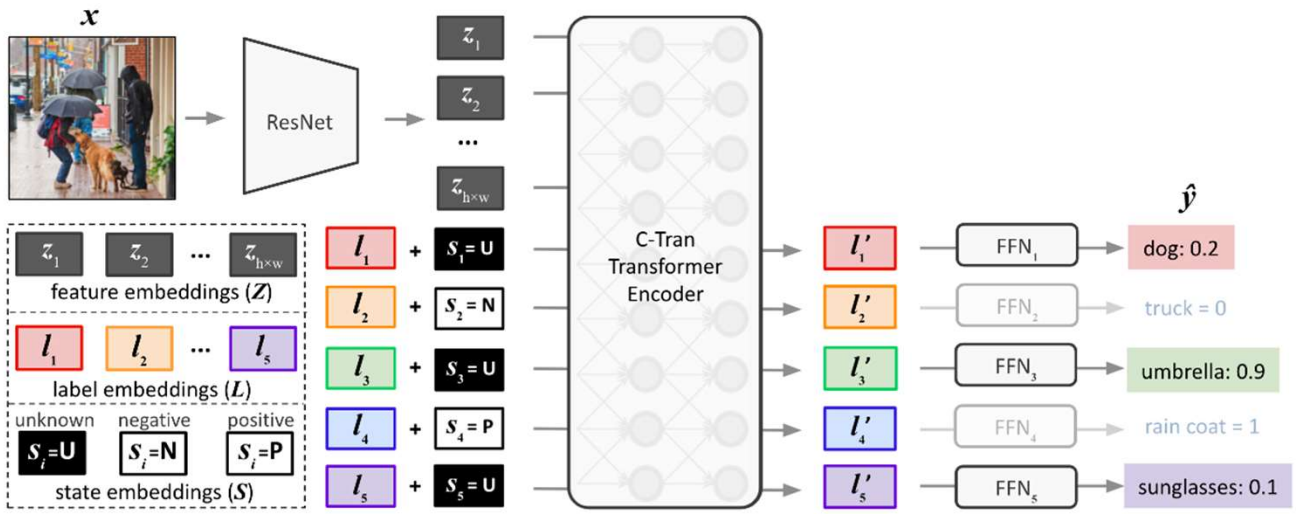
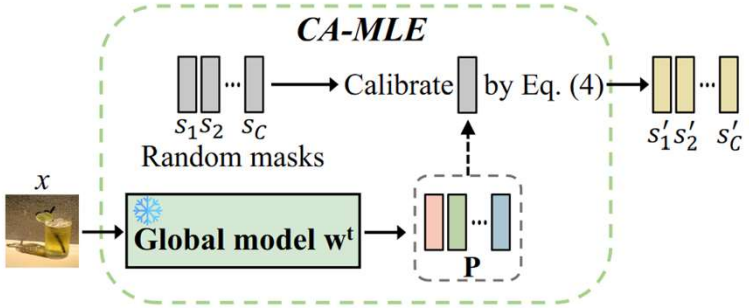


Figure 3. C-Tran architecture and illustration of label mask training for general multi-label image classification. In this training image, the labels *person*, *umbrella*, and *sunglasses* were randomly masked out and used as the unknown labels, $\mathbf{y}_u$. The labels *rain coat* and *truck* are used as the known labels, $\mathbf{y}_k$. Each unknown label is added the unknown state embedding U, and each known label is added its corresponding state embedding: negative (N) , or positive (P). The loss function is only computed on the unknown label predictions $\hat{\mathbf{y}}_u$.



CA-MLE generates the prediction by global model and calibrates the state embeddings.

$$s'_c : \begin{cases} unknown, & \tau - \varepsilon \le p_c \le \tau + \varepsilon \\ s_c, & \text{otherwise} \end{cases}$$

around 0.5 instead of being close to 1 or 0

Language-Guided Transformer for Federated Multi-Label Classification

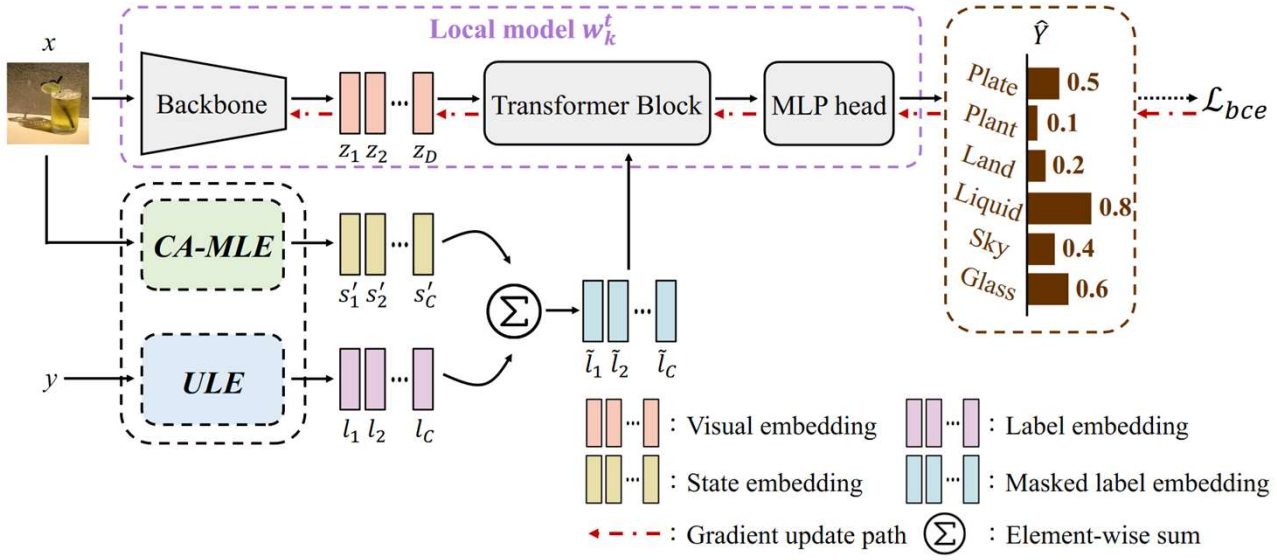
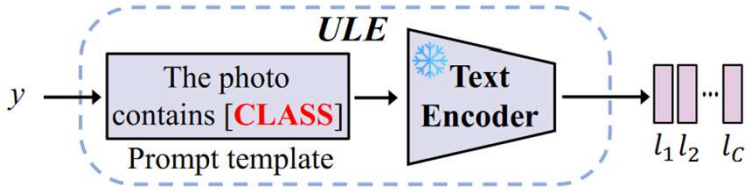


Figure 2: Overview of FedLGT. Given an image with multi-labels to predict, the global model from each communication round updates the local model with Client-Aware Masked Label Embedding (CA-MLE), which exploits partial label correlation observed at each client. In order to properly align local models for multi-label FL, universal label embeddings (ULE) are utilized in FedLGT. (Best viewed in color.)

$$w^{t+1} = \Sigma_{k=1}^{K} \frac{|D_k|}{|D|} w_k^t,$$



ULE advances the pre-trained label embeddings from CLIP for model aggregation purposes.

Performance Results

Metrics	C-AP	C-P	C-R	C-F1	O-AP	O-P	O-R	O-F1
Centralized (upper bound)								
ResNet	67.71	75.71	55.42	64.00	90.40	84.09	78.96	81.44
C-Tran	71.60	76.30	62.00	68.40	91.50	84.40	80.70	82.50
Federated								
FedAvg	40.63	42.74	26.53	32.74	77.39	78.35	56.31	65.53
FedC-Tran	56.00	49.40	38.20	43.10	88.10	83.10	72.50	77.40
Ours	**60.90**	**67.80**	**46.50**	**55.10**	**88.70**	**84.00**	**75.90**	**79.70**

Table 1: Comparisons of coarse-grained multi-label classification task on FLAIR. Bold denotes the best result under the FL setting.

Metrics	C-AP	C-P	C-R	C-F1	O-AP	O-P	O-R	O-F1
Centralized (upper bound)								
ResNet	20.26	32.97	10.92	16.40	47.95	68.73	30.04	41.81
C-Tran	27.50	33.10	13.30	18.90	54.20	71.00	34.70	46.60
Federated								
FedAvg	2.03	1.99	0.40	0.66	27.31	65.47	10.50	18.10
FedC-Tran	3.30	3.00	1.00	1.50	36.70	69.10	20.60	31.70
Ours	**10.60**	**6.50**	**1.40**	**2.30**	**42.20**	**69.80**	**21.90**	**33.40**

Table 2: Comparisons of fine-grained multi-label classification task on FLAIR. Bold denotes the best result under the FL setting.

Metrics	C-AP	C-P	C-R	C-F1	O-AP	O-P	O-R	O-F1
MS-COCO								
FedAvg	69.20	71.00	60.30	65.20	77.80	75.80	65.30	70.20
FedC-Tran	76.70	76.00	67.10	71.20	83.90	79.40	71.60	75.30
Ours	**78.30**	**77.20**	**70.00**	**73.40**	**84.70**	**80.20**	**73.70**	**76.80**
PASCAL VOC								
FedAvg	87.50	87.90	73.30	79.90	91.80	91.70	78.30	84.50
FedC-Tran	89.60	88.20	79.60	83.60	93.70	91.70	83.40	87.30
Ours	**90.80**	**88.80**	**82.50**	**85.50**	**94.10**	**91.80**	**85.30**	**88.40**

Table 3: Comparisons on MS-COCO and PASCAL VOC for our FedLGT with FL baselines.

Ablation Studies

Metrics	C-AP	C-F1	O-AP	O-F1
FedC-Tran	56.00	43.10	88.10	77.90
FedC-Tran + *CA-MLE*	56.10	45.00	88.30	78.40
FedC-Tran + *ULE*	59.70	54.90	88.30	78.90
Ours	**60.90**	**55.10**	**88.70**	**79.70**

Table 4: Abaltion studies of our FedLGT using coarse-grained task on FLAIR. Note that CA-MLE means client-aware masked label embedding, while ULE is universal label embedding. Bold denotes the best result.

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

THANKS