Language-Guided Transformer for Federated Multi-Label Classification

I-Jieh Liu[1], **Ci-Siang Lin**[1,2], **Fu-En Yang**[1,2], **Yu-Chiang Frank Wang**[1,2]

[1]Graduate Institute of Communication Engineering, National Taiwan University
[2]NVIDIA
{r11942087, d08942011, f07942077}@ntu.edu.tw frankwang@nvidia.com

Abstract

Federated Learning (FL) is an emerging paradigm that enables multiple users to collaboratively train a robust model in a privacy-preserving manner without sharing their private data. Most existing approaches of FL only consider traditional single-label image classification, ignoring the impact when transferring the task to multi-label image classification. Nevertheless, it is still challenging for FL to deal with user heterogeneity in their local data distribution in the real-world FL scenario, and this issue becomes even more severe in multi-label image classification. Inspired by the recent success of Transformers in centralized settings, we propose a novel FL framework for multi-label classification. Since partial label correlation may be observed by local clients during training, direct aggregation of locally updated models would not produce satisfactory performances. Thus, we propose a novel FL framework of **L**anguage-**G**uided **T**ransformer (**FedLGT**) to tackle this challenging task, which aims to exploit and transfer knowledge across different clients for learning a robust global model. Through extensive experiments on various multi-label datasets (e.g., FLAIR, MS-COCO, etc.), we show that our FedLGT is able to achieve satisfactory performance and outperforms standard FL techniques under multi-label FL scenarios. Code is available at https://github.com/Jack24658735/FedLGT.

1 Introduction

Federated Learning (FL) is a machine learning paradigm that enables multiple clients to collaboratively perform model training by leveraging their decentralized data. This approach offers significant advantages over traditional centralized approaches, as it mitigates the risks of privacy disclosure and is efficient for large-scale machine learning tasks without collecting large amounts of data on the server. FL has been successfully applied in real-world studies including health care, finance, and recommendation systems. As a pioneering work, the vanilla FL algorithm FedAvg (McMahan et al. 2017) learns from multiple local clients in a privacy-preserving manner, and the trained models or gradients are sent to the central server. Then, the server performed an aggregation step to obtain a global model. However, user data distribution often has non-IID (non-independent and identical distribution) characteristics in the real world, which is referred to as *data heterogeneity*. Data heterogeneity across different clients can be caused by various factors, such as demographic diversity, data collection methods, and sensor differences. In the presence of data heterogeneity, FedAvg may struggle to effectively leverage the data diversity across clients, resulting in degraded performance of the aggregated global model.

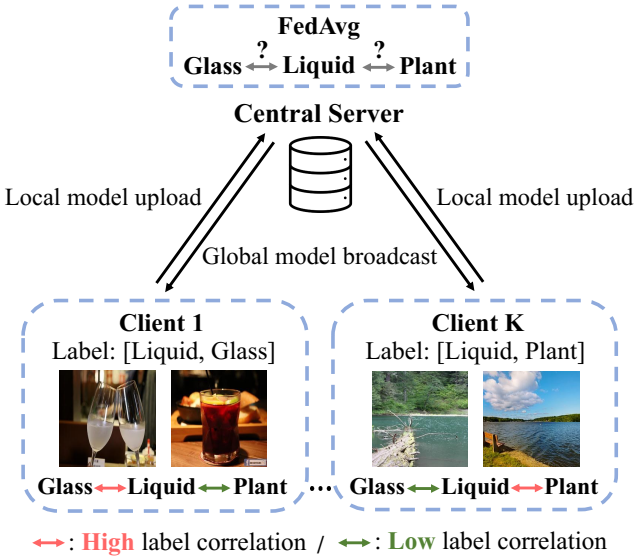

Figure 1: Challenges in multi-label federated learning. Since diverse label correlations are observed across clients, aggregating local models might not be sufficiently generalizable.

Typically, data heterogeneity contains the aspects such as label distribution skew or domain shift, presenting substantial obstacles in the progress of FL development. For instance, consider a scene understanding task in FL. In this scenario, certain clients possess a larger volume of indoor scene data, while others may have more outdoor scene data. As a result, their local models tend to be biased towards their respective scenes due to environmental variations in lighting, texture, etc., posing challenges in achieving a consensus during the model aggregation process (i.e., hard to obtain a robust global model). In addition, some clients may have more data for specific classes, resulting in a class im-

balanced issue that can also cause the performance degradation of the aggregated global model. Prior works (Li et al. 2020; Karimireddy et al. 2020; Gao et al. 2022; Li, He, and Song 2021; Li et al. 2021) tackle data heterogeneity mainly on standard single-label image classification benchmarking datasets, e.g., EMNIST (Cohen et al. 2017), CIFAR-10 (Krizhevsky, Hinton et al. 2009), etc. However, multi-label image classification is a more practical and challenging setting. For example, single-label image classification usually needs the model to recognize only one object or one main concept in an input image, whereas multi-label image classification aims to recognize all the object categories or concepts, which is more challenging due to the understanding of the inter-class relationships.

To mitigate performance degradation induced by data heterogeneity (e.g., label distribution skew) on FL, various works (Li et al. 2020; Karimireddy et al. 2020; Gao et al. 2022; Li, He, and Song 2021; Li et al. 2021; Huang, Ye, and Du 2022; Su et al. 2022; Mendieta et al. 2022; Wang et al. 2020; Tan et al. 2021; Zhuang et al. 2021; Luo et al. 2021) have been proposed to address this impact. For example, FedProx (Li et al. 2020) introduces a regularization term to mitigate the label distribution skew issues from the local learning step of FL. On the other hand, a previous FL method for addressing domain shift called FedBN (Li et al. 2021) proposes utilizing the batch normalization layer (Ioffe and Szegedy 2015) in local clients to capture the domain-specific information. However, these FL previous works neglect the emerging form of data heterogeneity such as multi-label issues but only focus on traditional single-label non-IID issues, leaving the *"multi-label FL"* task to remain unresolved and challenging.

Intuitively, it is achievable to directly apply the current multi-label classification frameworks from centralized learning (Chen et al. 2019b,a, 2020; Lanchantin et al. 2021) to FL. However, when performing multi-label classification under FL, the objects or concepts present in an image may exhibit significant variations across clients, potentially causing performance deterioration. For example, as Figure 1 depicted, the positive label set of each client is inconsistent or even disjoint, learning of the label correlation may have unfavorable impacts and degrade the global model performance after the aggregation step. Thus, one of the main challenges for multi-label image classification in FL is how to effectively learn label relationships, as well as precisely capture the complex connections between visual features and associated labels in a *privacy-preserving* manner.

In this paper, we propose a framework called *FedLGT* that aims to tackle the multi-label FL issues and verify our effectiveness on a challenging multi-label FL dataset (Song, Granqvist, and Talwar 2022) called FLAIR. Our FedLGT utilizes a pre-trained off-the-shelf text encoder by CLIP (Radford et al. 2021) to construct *universal label embedding*. The label embedding contains rich and distinct relationships among the labels. Utilizing this technique enables the clients to perform local learning with more discriminative label relationships. Besides, with the goal of training a robust global model from decentralized data, we also design a knowledge-transfer approach called *Client-Aware Masked Label Embedding* inspired by (Lanchantin et al. 2021) to assist local learning via knowledge from the global model. Specifically, the approach aims to transfer and distill the knowledge from the global model by encouraging the local models to learn more for the classes that the global model still predicts with relatively low confidence. Through conducting extensive experiments on FLAIR, our proposed framework is able to derive a global model with better generalization performance without any communication overheads.

Our contributions to this work are highlighted as follows:

- To the best of our knowledge, we are the first to tackle the problem of label discrepancy across different clients for multi-label FL.
- We propose *Client-Aware Masked Label Embedding* when training FedLGT. It is served as a customized model update technique while exploiting the label correlation at each client.
- We utilize *Universal Label Embedding* in FedLGT, which advances pre-trained label embedding derived from large-scale vision and language models (e.g., CLIP) for aligning local models for multi-label FL.

2 Related Works

2.1 Federated Learning

In general, the common vanilla FL algorithm (i.e., FedAvg (McMahan et al. 2017)) consists of several steps including local training, uploading client models, performing model aggregation, and broadcasting back to the clients on the server side. Under such FedAvg training procedures, the inconsistency between local objectives and the global objective would be more serious under data heterogeneity (e.g., domain shift). The main direction of existing previous methods could be broadly divided into *label distribution skew* or *domain shift*. Label distribution skew can arise due to disparate local data in different clients, leading to bias local models towards majority classes. On the other hand, domain shift does not specifically highlight the differences in label distributions across clients. Instead, the image features may exhibit significant variations, such as cartoons, sketches, etc., even under the same label annotation. In the subsequent sections, we mainly introduce the works that are popular or more relevant to our paper.

Label Distribution Skew Label distribution skew is a kind of common challenge in FL. For example, assume that some hospitals want to train a model to classify medical images. However, one client may have some rare diseases, while other clients have other common diseases. Thus, the label distribution will differ significantly or even disjoint (i.e., label spaces are not overlapped), leading to degradation of model performance. In terms of label distribution skew, existing FL methods focus on handling client local bias compared to the global model. For instance, FedProx (Li et al. 2020) restricts the gradient updates by introducing a proximal term to improve the convergence speed. SCAFFOLD (Karimireddy et al. 2020) designs a new control variate for each client by measuring the gradient dissimilarity

between local clients and the global model, utilizing this mechanism to refine the local clients' drift. More recently, FedDC (Gao et al. 2022) learns an auxiliary local drift variable by Expectation-Maximum (EM) algorithm to track the inconsistency of local-global models.

Domain Shift Domain shift (i.e., distribution shift) is another challenge among many real-world FL applications. For instance, a healthcare organization wants to perform training on patient data of certain diseases (i.e., the label distributions are the same) from different hospitals. Each of them maintains its own private dataset of medical images which may have different characteristics due to variations in the data acquisition process, patient populations, etc., resulting in potential domain shifts. Several previous works have been proposed to mitigate the domain shift of FL (Li et al. 2021; Huang, Ye, and Du 2022; Su et al. 2022). FedBN (Li et al. 2021) does not aggregate the parameters of batch normalization (BN) layers, so the domain-specific information can be preserved in local clients. FCCL (Huang, Ye, and Du 2022) builds a cross-correlation matrix on the server with the help of unlabeled public data. Besides, it balances the inter-domain and intra-domain information in the local training stage to tackle the domain shift. Recent FL works (Su et al. 2022; Guo et al. 2023; Chen et al. 2022; Sun et al. 2022) utilize vision-language pre-trained models, which perform prompt tuning during training rounds and thus reduce the number of learnable parameters. FedAPT (Su et al. 2022) learns a global adaptive network with the global prompt under the FL setting, then the framework generates domain-specific prompts for CLIP to handle the domain shift issue under FL. Inspired by this trend, our approach leverages vision-language pre-trained models (e.g., CLIP) to circumvent the challenges of training components that could deteriorate the model performance under the constraints of FL.

2.2 Multi-Label Image Classification

Centralized Learning Multi-label image classification has become an emerging research field since several realistic applications may view this task as a foundation, such as weakly supervised segmentation, scene understanding, etc. Since multi-label image classification requires recognition of all objects or concepts present in an image, it is critical yet challenging to understand the inherent relationships between different classes. There are many directions for multi-label image classification, such as improving loss functions (Ridnik et al. 2021), modeling label relationships (Chen et al. 2019b; Lanchantin et al. 2021), etc. The most related to our work is how to model the label relationships so that co-occurrence dependencies between different classes would be considered appropriately. Specifically, recent methods (Chen et al. 2019b,a, 2020) utilize graph formulation to represent label relationships. For example, ML-GCN (Chen et al. 2019b) build *Graph Convolutional Network (GCN)* to represent the correlation between objects or concepts, and such a method usually builds label correlation graphs based on extra knowledge from label co-occurrence statistics.

Based on Transformer, C-Tran (Lanchantin et al. 2021) fuses image features and the associated labels by *label mask training*. Similar to BERT (Devlin et al. 2018), C-Tran aims to capture semantic information by predicting the unknown labels on the images with the help of known labels. However, in FL settings, label co-occurrence information might differ across distinct clients. As confirmed later by our experiments, applying standard FL techniques with C-Tran might not be preferable.

Federated Learning Intuitively, one can utilize centralized multi-label learning models in FL schemes. That is, one can view each local client as a multi-label classification task and perform aggregation on the server side. However, existing centralized approaches require interaction with training data to capture a global view of the label relationships, which is infeasible under FL scenarios. For example in Fig.1, one client primarily captures images of drinks and glass, while the other client mainly takes pictures of landscapes and plants. Thus, the former client's model learns a high correlation between "liquid" and "glass", whereas the latter client's model learns a high correlation between "liquid" and "plant". However, aggregating these models on the server side could lead to confusion between these categories (i.e., "liquid", "glass", and "plant") for the global model, resulting in a degradation of performance.

To the best of our knowledge, recent FL works did not address multi-label learning tasks, and thus the above issue would limit the model performances. As discussed in the following section, we propose a *Transformer-based* model while tackling the inherent label distribution differences between clients. Compared to popular FL techniques like FedAvg, our method achieves impressive results on a large-scale dataset for multi-label FL (i.e., FLAIR (Song, Granqvist, and Talwar 2022)).

3 Proposed Method

3.1 Problem Formulation and Setup

In this work, we assume that K clients would be involved in the federated learning process in each communication round, and local private datasets are available for clients $D = \{D_1, D_2, ..., D_K\}$ during training. For each client, it can be viewed as solving a standard multi-label image classification task. To be more specific, given an image x with the corresponding label $y = [y_1, ..., y_C]$, where C is the number of classes, $y_i = 1$ indicates i-th class is present in the image, and $y_i = 0$ is not present, then the local clients attempt to predict the existence of each category of the image. The goal is to obtain a global aggregated model handling multi-label classification on FL that solves the objective

$$w^* = \arg\min_w \Sigma_{k=1}^{K} \frac{|D_k|}{|D|} \mathcal{L}_k(w), \tag{1}$$

In multi-label FL, each client and server share the same label space with a total of C categories. However, label distributions might differ across different clients. Thus, based on a recent SOTA centralized multi-label learning model of C-Tran (Lanchantin et al. 2021), our goal is to tackle the multi-label FL task.

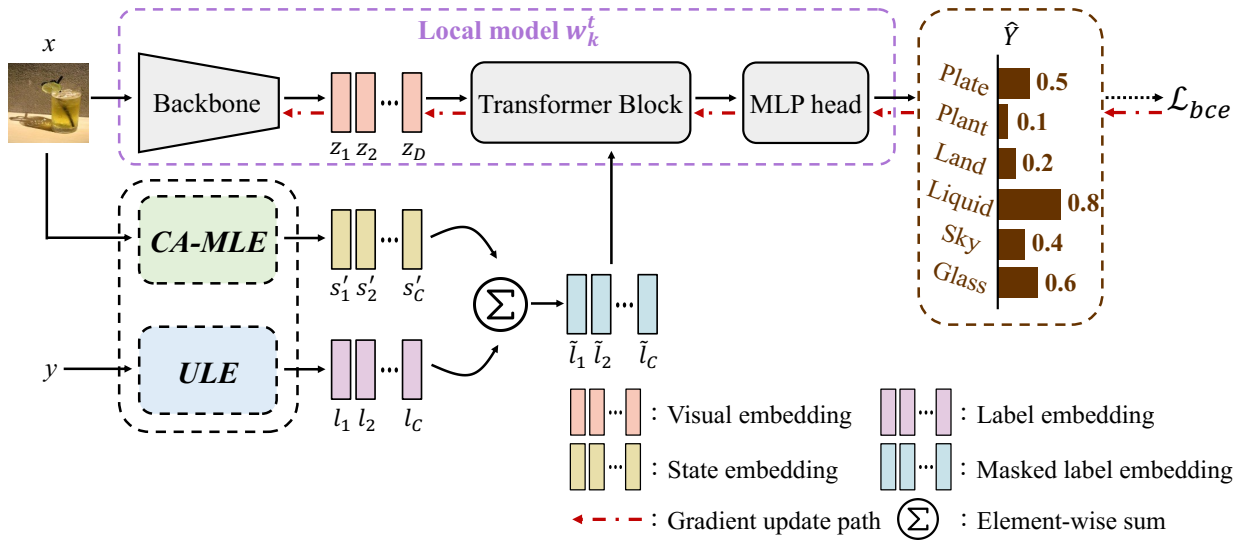

Figure 2: Overview of FedLGT. Given an image with multi-labels to predict, the global model from each communication round updates the local model with Client-Aware Masked Label Embedding (CA-MLE), which exploits partial label correlation observed at each client. In order to properly align local models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

3.2 A Brief Review for C-tran

C-Tran (Lanchantin et al. 2021) is a centralized *Transformer-based* (Vaswani et al. 2017) multi-label image classification framework, which is designed to observe image features and label correlations simultaneously. In C-Tran, the image features are extracted by ResNet (He et al. 2016), while the labels are described by both label L and state S embeddings. The label embeddings are defined as $L = \{l_1, l_2, ..., l_C\}$, where each $l_c \in \mathbb{R}^d$ represents c-th class labels (d denotes the embedding dimension). On the other hand, a set of state embeddings $S = \{s_1, s_2, ..., s_C\}$ (with each $s_c \in \mathbb{R}^d$) are viewed as tokens, indicating the presence of the corresponding labels of *unknown*, *positive*, and *negative*. Note the encoded token value for the associated state embeddings is -1, 1, and 0, respectively. Also, only the *unknown* state would contribute the loss to the model during training. With label and state embeddings, C-Tran proposes a training pipeline of *Label Mask Training (LMT)* that randomly masks partial amounts of labels and has the model perform prediction, implicitly exploiting label correlations. Specifically, the state embeddings would be added to the aforementioned label embeddings to form the masked label embeddings $\tilde{l_c}$ for LMT:

$$\tilde{l_c} = l_c + s_c, \tag{2}$$

where s_c is one of the states among *unknown*, *positive*, *negative*. Thus, the masked label embeddings can be formulated as $\tilde{L} = \{\tilde{l}_1, \tilde{l}_2, ..., \tilde{l}_C\}$ in the LMT process.

As shown in Fig. 2, the image features Z extracted by the vision backbone (e.g., ResNet (He et al. 2016)) would be concatenated with the masked label embeddings $\tilde{L}$, and sent into the transformer model. Thus, one has

$$\hat{Y} = w(x, \tilde{L}), \tag{3}$$

where x denotes the input image, $\tilde{L}$ represents the masked label embeddings, and $\hat{Y}$ is the predicted logit (w denotes the network model). However, as noted in Section 2, C-Tran cannot easily preserve label co-occurrence across different clients in FL settings. Thus, how to extend C-Tran for FL multi-label classification remains a challenging task.

3.3 Federated Language-Guided Transformer

In multi-label FL, we aim to learn a global model that generalizes to different clients with potential label distribution skews and diverse label correlations. Instead of simply performing model aggregation (like FedAvg) with pure visual input which might lead to degraded multi-label classification performances, we extend C-Tran and propose a novel learning scheme of Federated Language-Guided Transformer (FedLGT). As depicted in Fig. 2, we introduce model updating and feature embedding learning schemes for FedLGT, as presented below.

Client-Aware Masked Label Embedding To train global and client models in multi-label FL settings, how to tackle the domain shift or label skews across clients while jointly exploiting inherent label co-occurrence information is a challenging problem. At the t-th training round, each client applies the global model w^t to output the prediction vector $P = \{p_1, p_2, ..., p_C\}$, where each of $p_c \in [0, 1]$ represents a probability indicating the presence of the associated label. However, since the global model would not generalize to each client during training rounds, one would expect some class labels to be with lower confidence (i.e., p_c is around 0.5 instead of being close to 1 or 0). In such cases, a client-aware training strategy is necessary for updating the associate local model and calibrating the state embeddings s_c.

To address this problem, we propose *client-aware masked*

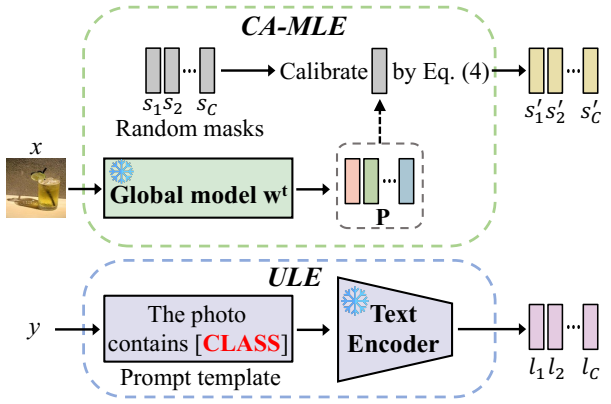

Figure 3: Details of CA-MLE and ULE. With the input image x and label y, CA-MLE generates the prediction by global model w^t and calibrates the state embeddings, while ULE advances the pre-trained label embeddings from CLIP for model aggregation purposes.

label embedding (CA-MLE) as depicted in Fig. 3 for local model updating. In CA-MLE, we would modify the original state embeddings of c-th class to be *unknown* if the p_c is uncertain, otherwise, the state embeddings remain unchanged. More precisely, the calibrated state embeddings $S' = \{s'_1, s'_2, ..., s'_C\}$ are defined as

$$s'_c : \begin{cases} unknown, & \tau - \varepsilon \leq p_c \leq \tau + \varepsilon \\ s_c, & \text{otherwise} \end{cases}, \tag{4}$$

where τ denotes the threshold (typically 0.5 as set in most multi-label works (Lanchantin et al. 2021; Liu et al. 2021)) to determine the presence of the corresponding class, and ε as the adaptable margin to describe the uncertainty. Note that only the unknown state would contribute losses to the training. Namely, if the probability lies in the interval, it indicates that the c-th class is uncertain for the global model w^t, and such a class is enforced to be further learned during that training round.

With calibrated state embeddings mentioned above, we view these embeddings as the masks to indicate which parts should be learned more during local training. We would not perform training by randomly generated masks as Eq. (2) stated. Instead, we utilize these calibrated state embeddings s'_c to each of the label embeddings to form the masked label embeddings to enhance the local training. Through the use of this calibration technique, we can harness the knowledge contained within the global model through state embeddings to guide the local training. Namely, it allows the local clients to focus more on those classes that are still challenging for the global model when local training, resulting in learning a better generalized and robust model after FL aggregation.

Universal Label Embedding In federated multi-label classification, the aggregated global model would not properly capture label correlation across different clients, since it does not have access to data for each client. Recall that, for each client, the local models may only capture the partial view of label correlations. Since an aggregation step is commonly deployed in FL to update the global model, it would be desirable to design a proper local model alignment mechanism, so that the inherent label co-occurrence information observed at each client would be properly shared and aggregated at the server.

To accomplish this, as depicted in Fig. 3, we propose to deploy *universal label embedding (ULE)* across clients in FedLGT. That is, we advance the large vision-language (VL) pre-trained model of CLIP (Radford et al. 2021) and utilize its (fixed) text encoder for deriving pre-aligned and fixed label embeddings. To be more specific, we build a prompt *"The photo contains [CLASS]"* as the input of the text encoder to generate the embedding for the associated c-th class label l_c. Such pre-trained/aligned label embeddings thus serve as the guidance during training of FedLGT, ensuring the updates of local models are aligned with predetermined embedding space and thus favoring the subsequent model aggregation process.

Algorithm 1: Training of FedLGT

Input: Communication rounds T, local epochs E, number of all clients K, the fraction of active client in each round ρ, the initial model w^0, k-th client dataset D_k;
Output: The trained global model w^t

Build universal label embedding L in Sec. 3.3
for $t = 1, 2, ..., T$ **do**
  Sample $R = \lceil \rho \cdot K \rceil$ clients to train
  for $k = 1, 2, ...R$ ***in parallel*** **do**
    $w_k^t \leftarrow$ `LocalUpdate` (w^t, L)
    $w^{t+1} \leftarrow \Sigma_{k=1}^{K} \frac{|D_k|}{|D|} w_k^t$
return w^t

`LocalUpdate` (w^t, L)
for $e = 1, 2, ..., E$ **do**
  $w_k^t \leftarrow w^t$
  for $(x_k, y_k) \in D_k$ **do**
    $P \leftarrow w^t(x_k, y_k, L)$
    Produce $\tilde{L}$ with P by **CA-MLE** in Sec. 3.3
    Update w_k^t by $\mathcal{L}_{bce}$
return w_k^t

3.4 Training of FedLGT

In FedLGT, we perform local model updates and global model aggregation in one communication round to achieve a federated learning process.

Local Model Update During the local model update, as depicted in Figure 2, each client k would go through CA-MLE with ULE to produce masked label embeddings. For CA-MLE, the client would receive an aggregated global model from the previous round, and it would make a prediction on a given image to obtain the logits. With these logits from the global model, it can be served as guidance for the

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.

local model in the training process. It implies that the local model w_k^t would identify the classes which are not properly described from the perspective of the global model. Thus, this allows the local model to focus more on those classes during training. As for ULE, the label correlations would be aligned according to fixed label embeddings across different clients. After passing these components, we follow Eq. (2) to perform element-wise addition of these embeddings to form **masked label embedding** to serve as the label inputs of the local model.

With masked label embedding, we leverage the self-attention mechanism in the transformer block to capture the relationships between the input image features and label correlations. To be more specific, we concatenate the masked label embeddings and image features extracted by the vision backbone (e.g., ResNet) as the input tokens of the transformer. Next, the output embeddings generated by the transformer are sent to the classifier (i.e., MLP head) to produce the multi-label classification results (i.e., $\hat{Y}$ in Figure 2).

As for the loss function to update the local model for client k (i.e., w_k^t shown in Figure 2), we adopt binary cross-entropy loss with our adjustment for masks, and only the masked labels would be optimized during training, which is formulated as:

$$\mathcal{L}_{bce} = -\sum_{\substack{s_c' \text{is } unknown \\ \forall c \in [1,C]}} y_c \log(p_c), \tag{5}$$

Global Model Aggregation Once the local training is done, the clients would upload their local models w_k^t to the server and perform the aggregation step in FL. Specifically, we utilize the vanilla FedAvg (McMahan et al. 2017) aggregation process, and it can be derived as follows:

$$w^{t+1} = \Sigma_{k=1}^{K} \frac{|D_k|}{|D|} w_k^t, \tag{6}$$

As soon as the training of our framework is converged, the trained aggregated global model is capable of performing multi-label image classification tasks by setting all the state embeddings to *unknown* for prediction.

With the above learning process, we are able to transfer the knowledge of the global model to local clients, with inherent and partial label correlation properly shared and updated. The pseudo-code of our proposed framework is described in Algorithm 1. Once the training of FedLGT is complete, we apply the learned model for testing.

4 Experiments

4.1 Experimental Setup

Datasets To demonstrate the effectiveness of our proposed learning framework, we conduct extensive evaluations on various benchmark datasets, including FLAIR (Song, Granqvist, and Talwar 2022), MS-COCO (Lin et al. 2014), and PASCAL VOC (Everingham et al. 2015). Specifically, we primarily evaluate our method on the recently introduced multi-label FL dataset, FLAIR. FLAIR is a large-scale multi-label FL dataset, which contains a wide variety of photos collected from real users on Flickr. FLAIR provides real-user data partitions with each input image in 256×256 pixels. Thus, FLAIR naturally captures various non-IID characteristics, including quantity skew (i.e., users have different numbers of samples), label distribution skew, and domain shift, leading to a more challenging scenario for FL. FLAIR is defined in a two-level hierarchy, one task called **coarse-grained** with 17 categories, the other called **fine-grained** with 1,628 categories. Besides, we also validate the performance of our method on centralized datasets partitioned artificially for FL, including MS-COCO and PASCAL VOC. To the best of our knowledge, these datasets have not been explored before in the context of multi-label FL. MS-COCO contains about 122,218 images containing common objects, and the standard multi-label formulation covers 80 object class annotations for each image. Moreover, the images in PASCAL VOC include multiple labels, corresponding to 20 object categories, which is a relatively easier task than FLAIR and MS-COCO.

Implementation Details We use ResNet-18 (He et al. 2016) as our vision backbone across all datasets mentioned above. For the universal label embeddings, we build a prompt *"The photo contains [CLASS]"* as the input of the text encoder for CLIP (Radford et al. 2021), whereas for the state embeddings, we build the prompt with one of the 'positive" or 'negative" for CLIP, and "unknown" is fixed with all zeros. Note that for FLAIR coarse-grained level since the categories are abstract, we build the prompts in two ways via the fine-grained level information. One collects the fine-grained labels for each coarse-grained label to build the prompts (i.e., [CLASS] parts), and the other builds all embeddings of the fine-grained labels, and for each coarse-grained label, taking average on corresponding fine-grained to build the coarse level embeddings. As for CA-MLE, we

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.

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.

set the threshold τ to 0.5 and the uncertainty margin ε is 0.02. For each round of local training, we train 5 epochs using the Adam (Kingma and Ba 2014) optimizer with a learning rate of 0.0001, and the batch size is set to 16. For the detail settings about FL, the communication round T is set to 50, and the fraction of active clients in each round is designed to achieve a level of participation equivalent to 50 clients, thus ensuring the data distribution is representative of the overall population. Besides, we observed that the statistics of FLAIR (Song, Granqvist, and Talwar 2022) suggest that the number of images across different users may have significant variations, leading to severe quantity skew issues. Thus, we follow some previous FL works (Li et al. 2019; Cho, Wang, and Joshi 2020) focusing on non-uniform client sampling to handle this issue. The concept is sampling clients at random such that the probability is the corresponding fraction of data at that client. With the help of this sampling scheme, we could guarantee our client quality, avoiding biased to very tiny local client datasets. Due to page limitations, we report the results with different sampling strategies in the *supplementary materials*. For all experiments, we implement our model by PyTorch and conduct training on a single NVIDIA RTX 3090Ti GPU with 24GB memory.

4.2 Performance Results and Analysis

As for the evaluation metrics, we follow the convention of works on multi-label image classification (Lanchantin et al. 2021; Ridnik et al. 2021) and use the metrics of per-class (C) and overall (O) average precision (i.e., C-AP and O-AP), precision (C-P, O-P), recall (C-R, O-R), F1 scores (C-F1, O-F1), with details provided in the *supplementary materials*. For the coarse-grained task in FLAIR, we first utilize C-Tran (Lanchantin et al. 2021) to conduct centralized learning experiments and compare the centralized results to pure ResNet-18 baseline. As shown in Table 1, the centralized C-Tran improves the performance. However, as we transfer the C-Tran directly to FL (i.e., FedC-Tran across all our tables), it surpasses the baseline FedAvg though. The gap between FedC-Tran and centralized is still present and has room for improvement. Hence, our proposed method could perform favorably against FedC-Tran (e.g., 4.9% on C-AP, 12% on C-F1, etc.). Similarly, as for the fine-grained task shown in Table 2, the centralized C-Tran still overcomes ResNet-18 baseline. But, FedC-Tran has a huge performance drop when transferred to FL compared to coarse-grained ones, which may be caused by the corrupted label relationships due to the aggregation step in FL. Thus, our method could avoid corrupt label relationships, leading to much more performance boosting (e.g., over $3\times$ on C-AP, $1.5\times$ on C-F1, etc.). Besides, as Table 3 presented, we also verify our effectiveness on extra datasets designed for centralized multi-label problems (MS-COCO, PASCAL VOC), and still achieve favorable performance over FedC-Tran baseline.

4.3 Ablation Studies

As reported in Table 4, we first perform ablation studies about the two components in our method. From "FedC-Tran + *CA-MLE*" of Table 4 (i.e., only using the client-aware masked label embedding), we observe that the improvement is not obvious (e.g., 0.1% on C-AP) compared to FedC-Tran baseline, which may be caused by the corrupted label correlations. Next, from "FedC-Tran + *ULE*" of Table 4 (i.e., only using universal label embedding), the improvement (e.g., 3.7% on C-AP) becomes more significant since the label embeddings are not corrupted by aggregation step. Thus, it can be seen the importance of ULE for the robustness and semantically label correlations. In the last row, our proposed method including the two aforementioned components is able to achieve satisfactory performance compared with the baselines (e.g., 4.9% on C-AP), and highlights the importance of ULE with the further improvements introduced by CA-MLE. Due to page limitation, additional ablation studies on the fine-grained FLAIR are provided in the *supplementary materials*.

5 Conclusion

In this paper, we tackle the challenging problems of multi-label FL. With the proposed framework of FedLGT, we are able to exploit local label correlation via learning client-aware masked label embedding. With universal label embedding derived from pre-trained vision and language model, the alignment of locally learned models can be performed in the same embedding space, allowing aggregation of such models for improved performance. We verified our proposed methods by conducting extensive experiments on the challenging dataset FLAIR as well as benchmarks of MS-COCO and PASCAL VOC. Our experiments confirmed the robustness and effectiveness of our proposed learning scheme for multi-label FL.

Acknowledgements

This work is supported in part by the National Science and Technology Council under grant 112-2634-F-002-007. We also thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

Chen, J.; Xu, W.; Guo, S.; Wang, J.; Zhang, J.; and Wang, H. 2022. FedTune: A Deep Dive into Efficient Federated Fine-Tuning with Pre-trained Transformers. *arXiv preprint arXiv:2211.08025*.

Chen, T.; Lin, L.; Chen, R.; Hui, X.; and Wu, H. 2020. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1371–1384.

Chen, T.; Xu, M.; Hui, X.; Wu, H.; and Lin, L. 2019a. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 522–531.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019b. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5177–5186.

Cho, Y. J.; Wang, J.; and Joshi, G. 2020. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*.

Cohen, G.; Afshar, S.; Tapson, J.; and Van Schaik, A. 2017. EMNIST: Extending MNIST to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, 2921–2926. IEEE.

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*.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Gao, L.; Fu, H.; Li, L.; Chen, Y.; Xu, M.; and Xu, C.-Z. 2022. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10112–10121.

Guo, T.; Guo, S.; Wang, J.; Tang, X.; and Xu, W. 2023. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*.

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.

Huang, W.; Ye, M.; and Du, B. 2022. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10143–10153.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. pmlr.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Lanchantin, J.; Wang, T.; Ordonez, V.; and Qi, Y. 2021. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16478–16488.

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. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2019. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; and Dou, Q. 2021. Fed{BN}: Federated Learning on Non-{IID} Features via Local Batch Normalization. In *International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Zhang, L.; Yang, X.; Su, H.; and Zhu, J. 2021. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*.

Luo, M.; Chen, F.; Hu, D.; Zhang, Y.; Liang, J.; and Feng, J. 2021. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34: 5972–5984.

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.

Mendieta, M.; Yang, T.; Wang, P.; Lee, M.; Ding, Z.; and Chen, C. 2022. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8397–8406.

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.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 82–91.

Song, C.; Granqvist, F.; and Talwar, K. 2022. FLAIR: Federated Learning Annotated Image Repository. *Advances in Neural Information Processing Systems*, 35: 37792–37805.

Su, S.; Yang, M.; Li, B.; and Xue, X. 2022. Cross-domain Federated Adaptive Prompt Tuning for CLIP. *arXiv preprint arXiv:2211.07864*.

Sun, G.; Mendieta, M.; Yang, T.; and Chen, C. 2022. Exploring Parameter-Efficient Fine-tuning for Improving Communication Efficiency in Federated Learning. *arXiv preprint arXiv:2210.01708*.

Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2021. Fedproto: Federated prototype learning over heterogeneous devices. *arXiv preprint arXiv:2105.00243*, 3.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33: 7611–7623.

Zhuang, W.; Gan, X.; Wen, Y.; Zhang, S.; and Yi, S. 2021. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4912–4921.