Breaking Barriers of System Heterogeneity: Straggler-Tolerant Multimodal Federated Learning via Knowledge Distillation

Jinqian Chen\textsuperscript{1,3}, Haoyu Tang\textsuperscript{1,}\textsuperscript{*}, Junhao Cheng\textsuperscript{1}, Ming Yan\textsuperscript{2}, Ji Zhang\textsuperscript{2}, Mingzhu Xu\textsuperscript{1}, Yupeng Hu\textsuperscript{1} and Liqiang Nie\textsuperscript{4}

\textsuperscript{1}School of Software, Shandong University
\textsuperscript{2}Alibaba Group
\textsuperscript{3}School of Software Engineering, Xi’an Jiaotong University
\textsuperscript{4}Harbin Institute of Technology (Shenzhen)

j1nqianchen6@gmail.com, \{tanghao258, lord.c, xumingzhu, huyupeng\}@sdu.edu.cn, \{ym119608, zj122146\}@alibaba-inc.com, nieliqiang@gmail.com

Abstract

Internet of Things (IoT) devices possess valuable yet private multimodal data, calling for a decentralized machine learning scheme. Though several multimodal federated learning (MFL) methods have been proposed, most of them merely overlook the system heterogeneity across IoT devices, resulting in the inadaptability to real-world applications. Aiming at this, we conduct theoretical analysis and exploration experiments on straggler impacts and uncover the fact that stragglers caused by system heterogeneity are fatal to MFL, resulting in catastrophic time overhead. Motivated by this, we propose a novel Multimodal Federated Learning with Accelerated Knowledge Distillation (MFL-AKD) framework, which is the first attempt to integrate knowledge distillation to combat stragglers in complex multimodal federated scenarios. Concretely, given the pretrained large-scale vision-language models deployed in the central server, we apply a fast knowledge transfer mechanism to conduct early training of local models with part of the local data. The early-trained model is then enhanced through the distillation of the pretrained large model and further trained on the remaining data. Extensive experiments on two datasets for video moment retrieval and two datasets for image-text retrieval demonstrate that our method achieves superior results with high straggler robustness.

1 Introduction

Nowadays, with the increase of multimedia data (e.g., images, videos, and texts) in daily life, the imperative to harness the wealth of multimedia data has become a hot topic, which has raised increasing research interests in many multimodal tasks such as text-image retrieval [Qu et al., 2021] and video moment retrieval [Gao et al., 2017; Wang et al., 2021]. The integration and storage of these data modalities across mobile and IoT systems present a significant challenge in training models with data from diverse devices.

Multimodal Federated Learning (MFL) [McMahan et al., 2017; Yang et al., 2020] has emerged as a promising approach to leverage multimodal data from various sources such as enormous IoT devices without privacy disclosure. These IoT devices always vary in computation capacity, communication bandwidth, energy power, and operation systems [Han et al., 2020]. All these factors lead to so-called system heterogeneity. However, existing MFL methods have overlooked this intrinsic property with a naive assumption: uniform capacities and the same model structures across all clients [Cheng et al., 2021; Wang et al., 2021]. Such an ideal assumption eliminates the heterogeneous nature of MFL, neglecting the severe time overhead and performance degradation caused by the stragglers.

Stragglers, or less efficient participants, have been recognized as a fundamental challenge in FL since its inception [Wu et al., 2020]. The huge extra time overhead with the performance degradation [Kairouz et al., 2021] caused by stragglers have led to the development of various mitigation methods, which can be broadly categorized into two groups: (1) employing relaxed synchronization, (2) improved scheduling and aggregation schemes. What’s worse, almost all these methods cannot be applied to MFL directly. For synchronization-based methods, the outdated models and dropped knowledge parts are deadly to multimodal scenarios. Additionally, aggregation-based methods face difficulties due to the expected diversity in model structures among clients, a result of the inherent system variability in MFL. This variation makes it challenging to directly combine local gradients, rendering the aggregation approach impractical in MFL.

Focusing on the issue, we carefully investigate the practical impact of stragglers in MFL under the real-world scenarios. Instead of simulation, we conduct all the explored experiments using real decentralized machines with different controlled computation capacities and communication bandwidth. As shown in Fig. 1, our findings reveal that the impact of stragglers in MFL is more severe than in horizontal single-modal FL, leading to significant delays and even training failure, and this problem is exacerbated as the number
The main contributions of this paper are summarized in experimental results validate the superior performance and the et al., 2021b; Hu et al., 2021] and video moment retrieval [Tang et al., 2020a]. However, these methods face challenges with communication bottlenecks and outdated local model updates during aggregation. Therefore, current research efforts primarily focus on straggler issues in synchronous FL. Common strategies involve setting fixed deadlines for all clients to update and share their local models, allowing clients to process data samples at varying speeds based on computational capacity [Smith et al., 2017; Li et al., 2020b; Han et al., 2020]. Additionally, [Reisizadeh et al., 2022] proposed to first train with faster clients and then gradually include stragglers towards the end of training. However, these approaches have not investigated the MFL scenario, which is more severely impacted by stragglers. They also lack a real-time update strategy in federated learning knowledge distillation environments, which is crucial for maintaining efficiency and effectiveness in MFL with multiple stragglers and extended delays.

2.2 Knowledge Distillation

Knowledge distillation (KD) aims to transfer the knowledge of a large and complex model (i.e., the teacher model) to a smaller and simpler model (i.e., the student model) so that a lightweight student with comparable performance to the teacher can be obtained [Hinton et al., 2015; Romero et al., 2015]. This technique often minimizes the discrepancy in probability distributions of teacher and student models between their final logical outputs [Hinton et al., 2015] or intermediate features [Komodakis and Pesquet, 2017].

Recently, knowledge distillation has demonstrated encouraging outcomes in compressing the client model size of unimodal FL [Huang et al., 2022]. For example, [Ahn et al., 2020] presented to transfer the knowledge from a global model to local models with a weighted loss function to balance the knowledge distillation and the federated learning objectives. [Li et al., 2020c] proposed a FedMD framework
that integrates transfer learning and knowledge distillation to facilitate federated learning in scenarios where individual clients possess their distinctive model designs.

Despite the progress they have made, directly extending those unimodal knowledge distillation methods into the multimodal FL scenarios will be highly inappropriate due to the heterogeneity of cross-modal data. More importantly, the straggler problem not addressed by these federated knowledge distillation methods will lead to more severe performance degradation and cost increases [Bonawitz et al., 2017].

2.3 Multimodal Federated Learning

Multimodal federated learning is an emerging area that involves using various data sources (e.g., text, image, and audio) from multiple clients to develop a multimodal model, ensuring data privacy across different clients. To address this issue, several methods have been proposed in recent years [Cheng et al., 2021; Wu et al., 2022; Li et al., 2020c; Zhao et al., 2021]. For example, [Liu et al., 2019] introduced FL to the vision-and-language grounding tasks and proposed to collaboratively extract diverse image representations derived from different tasks. Inspired by the idea of contrastive learning, [Huang et al., 2022] presented a CreamFL framework that regularizes local client training by incorporating both inter-modal and intra-modal contrasts to enhance the multimodal FL. Besides, [Wang et al., 2021] aimed at the video moment retrieval and proposed to attentively aggregate the model of grouped clients that are trained sequentially.

However, as these methods do not involve knowledge distillation and specialized acceleration strategy design, all of them will face severe performance degradation and communication delays when facing stragglers. In contrast, our MFL-AKD ensures its robustness to stragglers through the designed quick knowledge transfer strategy in KD process.

3 Revisiting Stragglers in Synchronized FL

To investigate the difference in the impact of stragglers in unimodal FL and multimodal FL, we here theoretically revisit and further conduct exploration experiments.

3.1 Problem Setup

Consider a multimodal federated learning scenarios with $K$ clients with up to $N$ modal data. Each client $c_i$ possesses $n_i$ samples, constituting the local dataset $\mathcal{D}_i = \{x_i^j, y_i^j\}^n_i$. The goal of multimodal federated learning is to collaboratively train a global model $f(\cdot; \theta_g)$ parameterized by $\theta_g$ by utilizing the decentralized datasets $\{\mathcal{D}_i\}^K$ in $T$ communication rounds. At the beginning of each round, the server will decide the participating clients set $\mathcal{P}$. Let $\mathcal{P}^r = \{\{c_i \in \mathcal{P}\}$.

The participant number in the $t$-th round is denoted as $\tau_t = |\mathcal{P}^t| = |\mathcal{P}|$. Each client $c_i \in \mathcal{P}$ is then required to conduct the training task on its local data $\mathcal{D}_i$. Considering the system heterogeneity across $\mathcal{P}$, there may exist $M^t$ stragglers in the $t$-th communication round, and $M^t \leq \tau_t$. Unimodal FL is the special case when $N = 1$ under this definition.

3.2 The Impact of Stragglers

System heterogeneity is an inevitable issue that occurs in the practical application of FL. Due to the difference in hardware, e.g., computing chips, battery power, and communication bandwidth, the ability to execute training tasks varies across different clients, which causes the well-known stragglers problem. Though numerous methods have been elaborated to deal with such an issue, the theoretical analysis of straggler impact in FL is yet to be established.

**Definition 1 (Time Consumption of Synchronized FL)** Consider a synchronized federated framework with $K$ clients $\{c_i\}^K_{i=1}$ for $R$ communication rounds, the server requires each participated client $c_i \in \mathcal{P}^r$ to conduct the training task on its local dataset $\mathcal{D}_i$. Let $\epsilon_i^r$ denote the total floating point operation number of client $c_i$ in the $r$-th round, $\gamma_i^r$ denote the comprehensive computation capacity measured by floating point operations per second (FLOPS). The time consumption of FL is defined as:

$$\Upsilon = \sum_{r=1}^{R} \max \left\{ \left( \frac{c_i^r}{\gamma_i^r} + \zeta_i^r \right) | i \in \mathcal{P}^r \right\}$$  \hspace{1cm} (1)

where $\max()$ is a function to extract the max element in a set and $\zeta_i^r$ is the communication time cost of client $c_i$ in the $r$-th round.

To further facilitate the numerical analysis, we introduce the sum of the averaged wasted time across all clients during a full federated training procedure as a metric to measure the absolute impact of stragglers in FL defined as below:

**Definition 2 (Averaged Wasted Time of Stragglers)** Given a federated framework with its relevant variables as introduced before, the averaged wasted time of stragglers is defined as:

$$\delta = \sum_{r=1}^{R} \sum_{i \in \mathcal{P}^r} \frac{1}{|\mathcal{P}^r|} \left| \frac{c_i^r}{\gamma_i^r} + \zeta_i^r - v \right|$$  \hspace{1cm} (2)

where $v = \max \left\{ \left( \frac{c_i^r}{\gamma_i^r} + \zeta_i^r \right) | i \in \mathcal{P}^r \right\}$ is the time consumption of the lowest client in $r$-th round.

We now claim our analysis of straggler impacts as indicated in Proposition 1.

**Proposition 1** The impact of stragglers is correlated with: (1) Task Difficulty $\Phi$; (2) Balance of dataload on unit computation capacity; (3) Communication consumption.

**Analysis** (1) Task difficulty $\Phi$ closely relates to the required communication rounds $R$ for desired model performance. The convergence speed within a federated algorithm is influenced by the convexity and smoothness of the objective functions. Therefore, simpler tasks that satisfy the conditions of convexity and smoothness, tend to converge more rapidly. (2) Workload on unit computation capacity is decided by the data amount and the model capacity. The more imbalance the dataload on unit computation, the more gap exists in the time consumption across all clients.

**Discussion** (1) The common trick to mitigate severe stragglers involves setting a response threshold, effectively reducing $v$ to lessen the time consumption gap $\left| \frac{c_i^r}{\gamma_i^r} + \zeta_i^r - v \right|$. (2) The communication budget is generally constant, as the uploaded content consistently includes the gradient or parameters of the current local model. Therefore, communication
consumption $\zeta_i$ mainly depends on the communication bandwidth, which is hard to manually optimize in federated algorithm design.

Comparison of multimodal FL with unimodal FL: (1) Tasks in multimodal FL are often more difficult than those in unimodal FL, as they involve the joint relationships between different modalities. (2) Workload is more imbalanced across clients in multimodal FL than in unimodal FL, as data in different modalities possesses different data structures with different storage requirements. Those differences magnify the impact of stragglers in multimodal FL, calling for the establishment of a straggler-robust multimodal FL framework.

4 The Proposed Method

In this section, we introduce our proposed straggler-robust multimodal federated framework MFL-AKD. We choose the vision-language tasks as representative multimodal tasks. Specifically, we consider text-to-image and video moment retrieval tasks, which aim to match a given sentence $x$ with its most relevant image or video moment $y$.

4.1 Motivation

We now briefly revisit our motivation for the proposal of MFL-AKD. As described before, system heterogeneity is an inevitable issue in the practical application of multimodal federated learning, causing stragglers in each communication round. However, such an issue has been largely overlooked with a naive conjecture that the impact of stragglers in MFL is similar to that in the unimodal FL. With the analysis proposed in Section 3.2, the impact of stragglers in multimodal FL is much more severe than that in unimodal FL, which calls for the establishment of a straggler-robust multimodal federated framework. To develop such a framework, the following requirements should be satisfied: From the system perspective: (1) Lightweight but effective client-deployed model to alleviate the workload of clients. (2) Powerful and large server-deployed model to sufficiently handle the difficult multimodal tasks. From the scheduling perspective: (1) Balanced workload decided by computation capacities across all clients; (2) Efficient aggregation schema to accelerate the convergence.

4.2 Framework Overview

In response to these requirements, we develop a straggler robust multimodal federated method named MFL-AKD. To handle the challenge brought by multimodal tasks, MFL-AKD deploys a powerful pretrained vision-language model CLIP $F(\cdot; \theta_C)$ on the server and allows clients to use its desired model $f_i(\cdot; \theta_i)$ in arbitrary structures. To facilitate the convergence of the federated algorithm and allow the dynamic workload to prevent stragglers, we have designed a novel Fast Knowledge Transfer Mechanism (FKTM) in MFL-AKD. FKTM requires all clients to extract part of its local data to conduct workload-balanced early training before the ordinary training. Such a partial extraction ensures the balance of workload on computation capacity, largely reducing the wasted time of stragglers. The early-trained models are then uploaded to the server to conduct a warm-up utilizing the CLIP model through knowledge distillation. By doing this, FKTM not only mitigates the time consumption gap between clients but also accelerates the convergence of the federated framework. The application of KD also allows the deployment of models in different structures on clients. Note that we here mainly introduce MFL-AKD with the same model architectures deployed in clients to facilitate the comparison with other MFL methods. The overall framework of MFL-
AKD is illustrated in Figure 2.

Commencing the $r$-th round, the central server first identifies the participated client set $\mathbb{P}_r$. For each client $c_i \in \mathbb{P}_r$, the server distributes current global model parameter $\theta_g$ to initialize its local model. Concerning the possible imbalance workload on computation capacity $c_i$ among clients set $\mathbb{P}_r$, MFL-AKD requires each client $c_i \in \mathbb{P}_r$ to randomly extract a mini-batch of local data $D_i$ according to its computation capacity $c_i$ in current round and further conduct early-training to get the local model parameter $\theta_{el,i}^r$, which is further asynchronously uploaded to the server. Receiving the uploaded early-trained parameter $\theta_{el,i}^r$ of client $c_i$, the FKTM in the server then utilizes the CLIP model to conduct knowledge distillation to warm up the early-trained local model, accelerating the convergence of the federated frameworks. The distilled early-trained local model $\theta_{el,i}^r$ is then distributed to its corresponding client $c_i$ and further conduct model training on the rest data to get updated local model $\theta_i$. All the updated local model $\{\theta_i^r\}_{i \in \mathbb{P}_r}$ is then uploaded to the server to get global model $\theta_{g}^{r+1}$ for next round training.

4.3 Multimodal Local Training

Without loss of generality, we introduce the local training of MFL-AKD under the multimodal retrieval tasks. The input instance $\{x_i, y_i\}$ in the local dataset $D_i$ denotes a sentence $x_i$ and its related image or video moment $y_i$, and the training goal is to obtain a local multimodal encoder $f_i(\cdot; \theta_i)$ that characterizes the visual language well. Considering the excellence of transformer in cross-modal embedding, we adopted a transformer-structured encoder [Vaswani et al., 2017].

Specifically, $f_i$ encodes $x_i$ and $y_i$ into their respective embeddings $h_{x_i}$ and $h_{y_i}$. The local loss function that minimizes the distance between $h_{x_i}$ and $h_{y_i}$ is defined as:

$$
\min_{f_i} \frac{1}{n_i} \sum_{n=1}^{n_i} L(h_{x_i}, h_{y_i})
$$

where $n_i$ denotes the size of $D_i$, and $L$ denotes the mean square error.

4.4 Fast Knowledge Transfer Mechanism (FKTM) via Centralized Knowledge Distillation

We here specifically introduce the designed FKTM with a centralized knowledge distillation process in MFL-AKD. As analyzed before, the impact of stragglers on a particular federated framework is highly correlated with its convergence speed and balance of workload on unit computation capacity across clients. To effectively handle the straggler problem and prevent entirely dropped clients, FKTM requires the clients $c_i \in \mathbb{P}_r$ to early-train the initialized local model $\theta_{el,i}^r$ on a small fraction of its local data while remaining the workload balance. After receiving the local early-trained model $\theta_{el,i}^r$, FKTM further enhances the ability of the model by distilling knowledge from the powerful teacher model. Specifically, a large pretrained CLIP model $F(\cdot; \theta_C)$ is adopted as the teacher. For the input pair $(x_i, y_i)$, the CLIP teacher and the student encoder $\theta_{el,i}$ embed them into their corresponding representations and further conduct knowledge distillation through:

$$
L_{dist} = KL \left( F(x_i^r | \theta_C) || f(x_i^r | \theta_{el,i}) \right)
$$

The distilled local model $\theta_{el,i}^r$ of client $c_i$ is then redistributed to the client to continue the training on the rest of its local data. FKTM sufficiently utilizes the prior knowledge in pretrained large multimodal models to help the early-trained model rapidly adapt to its local multimodal data and get experience from the prior knowledge. Such a mechanism significantly accelerates the convergence of the MFL-AKD, making the framework tolerant to stragglers and leaving readiness models for each client to combat the potential drop-offs during the subsequent training.

5 Experiments

We evaluated our framework on two fundamental vision-language tasks: image-text retrieval on the Flickr30k and MS COCO datasets, and video moment retrieval on the Charades-STA and ActivityNet Captions datasets.

5.1 Experimental Setup

Datasets We use four popular multimodal datasets in text-image and text-video retrieval tasks. Details are given below:

1. Flickr30k [Young et al., 2014]: This image-text retrieval dataset consists of 31,784 images, each of which is manually annotated with five different sentence descriptions. As in [Qu et al., 2021], we adopt a transformer-structured encoder [Vaswani et al., 2017].

Specifically, $f_i$ encodes $x_i$ and $y_i$ into their respective embeddings $h_{x_i}$ and $h_{y_i}$. The local loss function that minimizes the distance between $h_{x_i}$ and $h_{y_i}$ is defined as:

$$
\min_{f_i} \frac{1}{n_i} \sum_{n=1}^{n_i} L(h_{x_i}, h_{y_i})
$$

where $n_i$ denotes the size of $D_i$, and $L$ denotes the mean square error.

4.4 Fast Knowledge Transfer Mechanism (FKTM) via Centralized Knowledge Distillation

We here specifically introduce the designed FKTM with a centralized knowledge distillation process in MFL-AKD. As analyzed before, the impact of stragglers on a particular federated framework is highly correlated with its convergence speed and balance of workload on unit computation capacity across clients. To effectively handle the straggler problem and prevent entirely dropped clients, FKTM requires the clients $c_i \in \mathbb{P}_r$ to early-train the initialized local model $\theta_{el,i}^r$ on a small fraction of its local data while remaining the workload balance. After receiving the local early-trained model $\theta_{el,i}^r$, FKTM further enhances the ability of the model by distilling knowledge from the powerful teacher model. Specifically, a large pretrained CLIP model $F(\cdot; \theta_C)$ is adopted as the teacher. For the input pair $(x_i, y_i)$, the CLIP teacher and the student encoder $\theta_{el,i}$ embed them into their corresponding representations and further conduct knowledge distillation through:

$$
L_{dist} = KL \left( F(x_i^r | \theta_C) || f(x_i^r | \theta_{el,i}) \right)
$$

The distilled local model $\theta_{el,i}^r$ of client $c_i$ is then redistributed to the client to continue the training on the rest of its local data. FKTM sufficiently utilizes the prior knowledge in pretrained large multimodal models to help the early-trained model rapidly adapt to its local multimodal data and get experience from the prior knowledge. Such a mechanism significantly accelerates the convergence of the MFL-AKD, making the framework tolerant to stragglers and leaving readiness models for each client to combat the potential drop-offs during the subsequent training.

5 Experiments

We evaluated our framework on two fundamental vision-language tasks: image-text retrieval on the Flickr30k and MS COCO datasets, and video moment retrieval on the Charades-STA and ActivityNet Captions datasets.

5.1 Experimental Setup

Datasets We use four popular multimodal datasets in text-image and text-video retrieval tasks. Details are given below:

1. Flickr30k [Young et al., 2014]: This image-text retrieval dataset consists of 31,784 images, each of which is manually annotated with five different sentence descriptions. As in [Qu et al., 2021], 29,784, 1,000, and 1,000 images with paired sentences are adopted for training, validation, and testing, respectively.

2. MSCOCO [Lin et al., 2014]: This image-text retrieval dataset contains 123,287 images, each of which is paired with five annotated sentences. For fair comparisons, the public dataset split is adopted [Qu et al., 2021], i.e., 113,287, 5000, and 5000 images for training, validation, and testing, respectively. Besides, the MSCOCO 5Fold 1K setting is adopted for evaluation, where the average results are over 5-fold of 1,000 testing images.

3. Charades-STA [Gao et al., 2017]: This dataset video moment retrieval is manually annotated by [Gao et al., 2017], which contains 6,672 videos with 29.76 seconds long on average. The number of sentence-video pairs is 16,127 in total. Following the common settings [Gao et al., 2017], we divide those pairs into two parts, i.e., 12408 pairs for training and 3720 pairs for testing, respectively.

4. ActivityNet Captions (Anet) [Krishna et al., 2017]: This dataset video moment retrieval dataset contains 14926 videos with an average duration of 120 seconds. The sentence-video pairs are 71,957 in total, where the corresponding sentences are longer with more complicated semantics. Following [Gao et al., 2017], we adopt 37,417, 17,505, and 17,031 sentence-video pairs for training and validation, and testing, respectively.

Evaluation metrics. The standard Recall@K (R@K for short) and R@n, IoU@m are adopted as the evaluation metrics for the image-text retrieval and video moment retrieval
### Table 1: Performance comparison of Text-Image Retrieval on Flickr30k and MSCOCO datasets (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>Flickr30k</th>
<th>MSCOCO 5Fold 1K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text Retrieval</td>
<td>Image Retrieval</td>
</tr>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>FedAvg [McMahan et al., 2017]</td>
<td>67.9</td>
<td>47.8</td>
</tr>
<tr>
<td>FedIoT [Zhao et al., 2021]</td>
<td>67.2</td>
<td>45.4</td>
</tr>
<tr>
<td>FedCG [Wu et al., 2022]</td>
<td>66.5</td>
<td>44.9</td>
</tr>
<tr>
<td>FedMD [Li et al., 2020c]</td>
<td>67.4</td>
<td>47.9</td>
</tr>
<tr>
<td>FedGEMS [Cheng et al., 2021]</td>
<td>69.4</td>
<td>50.9</td>
</tr>
<tr>
<td>FedVMR [Wang et al., 2021]</td>
<td>69.3</td>
<td>51.2</td>
</tr>
<tr>
<td>MFL-AKD (w/o. KD)</td>
<td>66.8</td>
<td>46.9</td>
</tr>
<tr>
<td>MFL-AKD (w/o. FKTM)</td>
<td>68.3</td>
<td>51.3</td>
</tr>
<tr>
<td>MFL-AKD</td>
<td>69.0</td>
<td>51.3</td>
</tr>
</tbody>
</table>

### Table 2: Performance comparison of Video Moment Retrieval on Charades-STA and Anet datasets (%).

<table>
<thead>
<tr>
<th>Models</th>
<th>Charades-STA</th>
<th>Anet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>FedAvg [McMahan et al., 2017]</td>
<td>50.5</td>
<td>82.4</td>
</tr>
<tr>
<td>FedIoT [Zhao et al., 2021]</td>
<td>47.9</td>
<td>80.4</td>
</tr>
<tr>
<td>FedCG [Wu et al., 2022]</td>
<td>39.4</td>
<td>78.9</td>
</tr>
<tr>
<td>FedMD [Li et al., 2020c]</td>
<td>51.7</td>
<td>82.9</td>
</tr>
<tr>
<td>FedGEMS [Cheng et al., 2021]</td>
<td>52.9</td>
<td>83.6</td>
</tr>
<tr>
<td>FedVMR [Wang et al., 2021]</td>
<td>53.0</td>
<td>84.7</td>
</tr>
<tr>
<td>MFL-AKD (w/o. KD)</td>
<td>49.3</td>
<td>79.2</td>
</tr>
<tr>
<td>MFL-AKD (w/o. FKTM)</td>
<td>49.2</td>
<td>80.1</td>
</tr>
<tr>
<td>MFL-AKD</td>
<td>53.1</td>
<td>84.6</td>
</tr>
</tbody>
</table>

5.2 Performance Comparison

We first compare the performance of MFL-AKD with SOTA multimodal FL methods on four multimodal datasets. Table 1 and Table 2 display the retrieval performance of MFL-AKD and other baselines for video moment retrieval and text-image retrieval tasks, respectively. Note that the best results are highlighted and the second ones are underlined.

From those results, the following observations stand out.
We further perform ablation studies to validate the effectiveness of the components of the MFL-AKD framework. The results are shown in Table 3 and Table 4, which demonstrate that the proposed MFL-AKD framework achieves superior performance compared to the strong baselines on various multimodal benchmarks.

For the text-image retrieval task, the proposed MFL-AKD model consistently surpasses all baselines across all reported metrics. As for the "R@1" metric on both datasets, the proposed MFL-AKD model achieves substantial improvements over the strongest FedVMR baseline. The knowledge distillation of pretrained model is now conducted on the global updated model after the aggregation of the clients’ model, which validates the effectiveness of key components of MFL-AKD.

5.3 Robustness on Stragglers

We further conduct experiments to validate the effectiveness of MFL-AKD in combating the stragglers in MFL scenarios. To measure the severity of the straggler problems, we adopt the same implementation strategy with the performance comparison experiments, while manually changing the straggling time. We use the time consumption $\tau$ to compare the robustness of stragglers among various methods. Table 3 and Table 4 report the experimental results on Flickr30k for image-text retrieval tasks and Anet for video moment retrieval tasks, respectively. As demonstrated in the table, MFL-AKD exhibits the shortest time consumption under different straggling times, and when the straggling time varies from the lowest to the highest, the time consumption only increases by around 5k for text-image retrieval and 10k for video moment retrieval, which is far lower than that of other baselines.

5.4 Ablation Studies

To evaluate the contribution of different components in our proposed MFL-AKD framework, we further conduct ablation experiments to validate the effectiveness of the design of MFL-AKD. Specifically, we gradually remove key components of our framework and obtain the following model variants.

- **MFL-AKD (w/o. FKTM):** We discard the FKTM from MFL-AKD, which directly returns the aggregated model of early-trained models without learning from pretrained large model.
- **MFL-AKD (w/o. KD):** We remove the knowledge distillation mechanism from MFL-AKD, which directly returns the aggregated model of early-trained models without learning from pretrained large model.

The results are illustrated in the corresponding Table 1, 2, 3, and 4 respectively. From the results, missing of KD or FKTM will lead to a significant loss of performance and robustness to stragglers, which validates the effectiveness of key components of MFL-AKD.

6 Conclusion

In this paper, we make the first attempt to analyze and combat stragglers in multimodal FL by integrating knowledge distillation, especially for vision-language scenarios. We propose a novel straggler-robust multimodal FL method named MFL-AKD. With the knowledge distilled from the vision-language pretrained model, the cross-modal semantic representations of the local model are greatly and rapidly enhanced, resulting in a remarkable improvement in convergence speed. Moreover, we also design the Fast Knowledge Transfer Mechanism to allow the balance of workload and enable early training of local models to handle the stragglers. Through experiments on video moment retrieval and text-image retrieval datasets, we have verified that our method can significantly alleviate the impact of stragglers while achieving remarkable retrieval accuracy.
Acknowledgments

This work was supported in part by the Alibaba Group through Alibaba Innovative Research Program, No.21169774; in part by the National Natural Science Foundation (NSF) of China, No.62206156, No.62276155, No.72004127, and No.62206157; in part by the NSF of Shandong Province, No.ZR2021MF040 and No.ZR2022QF047; in part by the Key R&D Program of Shandong Province, China (Major Scientific and Technological Innovation Projects), No.2022CXGC020107.

References


[Huang et al., 2022] Jie Huang, Zheng Liu, Dong Yang, and Yang Li. Multimodal federated learning via contrastive representation ensemble. In AAAI, 2022.


[Radford et al., 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-


[Zhao et al., 2021] Liang Zhao, Junlin Wang, Zhigang Chen, Qiang Wu, and Jie Chen. Multimodal federated learning on iot data. *IEEE Internet of Things Journal*, 2021.