An Empirical Study of Federated Learning and Video Representations for Human Action Recognition

by

Assanali Abu

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

May 2024

Author .....
Department of Computer Science
May 3, 2024

Certified by .....
Nguyen Anh Tu
Assistant Professor
Thesis Supervisor

Certified by .....
Min-Ho Lee
Assistant Professor
Thesis Supervisor

Accepted by .....
Yelyzaveta Arkhangelsky
Dean, School of Engineering and Digital Sciences

An Empirical Study of Federated Learning and Video Representations for Human Action Recognition

by

Assanali Abu

Submitted to the Department of Computer Science
on May 3, 2024, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

This thesis explores the integration of Federated Learning in Human Action Recognition, with a focus on both supervised learning and Few-Shot Learning methodologies using RGB, skeleton, or fusion data. By employing advanced Deep Learning models and leveraging large-scale, diverse datasets, we demonstrate significant advancements in HAR, crucial for enhancing smart video surveillance systems. Our novel FL framework addresses the challenges associated with centralized learning, such as substantial resource allocation and potential confidentiality violations, by training models in a decentralized manner. This approach enhances privacy and efficiency, utilizing diverse data across devices to improve generalizability. We present a comprehensive evaluation of various 3D-CNN and Transformer-based architectures, emphasizing the effects of pre-training and comparing FL algorithms, specifically FedAvg and FedProx, under realistic settings. Our findings reveal that integrating pre-trained 3D-CNNs and Transformers with optimal FL configurations can significantly enhance HAR performance, enabling our models to compete with and, in some cases, surpass centralized learning counterparts.

Thesis Supervisor: Nguyen Anh Tu
Title: Assistant Professor

Thesis Supervisor: Min-Ho Lee
Title: Assistant Professor

Acknowledgments

I express my sincere gratitude to Professor Anh Tu Nguyen for his invaluable guidance and support throughout my thesis. His expertise and encouragement have been pivotal to my research journey. Special thanks also to my labmates, Nursultan Makhanov and Nartay Aikyn, for their friendship, insightful discussions, and collaborative spirit, which have significantly enriched my experience and contributed to my work. I am deeply grateful to each of them for their role in my academic and personal growth.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Using state-of-the-art Deep Learning (DL) models and large-scale diverse datasets, Human Action Recognition (HAR) has achieved tremendous progress in recent years [22]. This advancement has played a pivotal role in the field of smart video surveillance systems, especially in enhancing the capability of these systems to analyze and classify human actions using CCTV cameras. The progress is crucial to ensure safety and protection in a variety of environments. Traditionally, the development of the robust video action recognition models requires a large amount of diverse and annotated video data samples and extensive computational resources. It is apparent from the current DL methods [5, 51], that the common approach of centralized learning involves significant resource allocation for storage and communication, as it demands the transmission of the local data into a centralized server rising deployment concerns with regards to the resource management. It also goes without saying that the risk of violation of confidentiality is being risen when the sensitive information such as personal identity and behavioural patterns are being stored in a central server without consent.

To address this issues, Federated Learning (FL) [25] offers a promising alternative that enables model training in a decentralized manner. FL enhances the privacy and efficiency, enabling the collaborative model training without a need to centralize sensitive data. In a nutshell, FL gathers and orchestrates the individual models computed on each device, training a globally shared model on a central server by aggregating

these local computations. This approach not only decreases the data leakage risks, but also utilizes diverse and varied data across different devices improving the system's ability to be generalized and recognize more complex patterns.

In the realm of Human Action Recognition, the integration of Federated Learning (FL) within supervised learning settings represents a significant stride towards realizing more privacy-preserving, efficient, and scalable HAR systems. Supervised learning, a foundational pillar in the development of HAR models, traditionally relies on extensive datasets that include a variety of data types such as RGB videos, skeleton data, or a fusion of both to train models capable of accurately recognizing human actions. However, the centralized aggregation of such diverse and potentially sensitive data raises substantial concerns over privacy, data security, and the logistical overhead associated with the storage and processing of massive datasets. Employing FL for supervised learning in HAR, using RGB, skeleton, or fusion data, opens new avenues for crafting state-of-the-art recognition systems that are not only effective across a broad spectrum of scenarios but are also aligned with stringent privacy requirements and operational efficiency.

Despite these advantages, both FL and traditional DL methods share the challenge of requiring large amounts of labeled data, which is costly and labor-intensive to produce. Most of the time, in order to train a robust video action recognition model, the task would require hundreds of samples per class [20]. The dynamic nature of human actions and the continuous emergence of novel action classes further complicate data collection and annotation efforts. Additionally, for end users or institutions such as airports, schools, banks, and hospitals, there is a challenge in collecting new video data samples for unfamiliar activity categories. Often, these new categories associated with emerging problems arise at a time when there is a problem in collecting adequate annotated data to effectively address these new problems. The task of preparing a new set of training data becomes especially difficult in the face of a vast variability in the human actions. Also, FL systems assume that there are sufficient number of data samples for each of the participant i.e. device, which is not often the case. In the real-world environment, the completeness of data across multiple devices vary,

and the data is often distributed unevenly. This results in two following concerns. First is the model bias. The FL model may learn more from data-rich participants, leading to biased predictions. Secondly, reduced generalizability. The model's ability to perform well on new or previously unknown data is reduced because it was not trained on a sufficiently diverse dataset.

Herein, Few-Shot Learning (FSL) [45] emerges as a solution, enabling models to recognize new action classes with minimal training samples. The common strategy for addressing Few-Shot Learning (FSL) involves meta-learning, which seeks to mimic human learning by applying knowledge from previous tasks to new, yet related ones. Specifically, this involves training on a variety of action videos from established classes and applying the learned meta-knowledge to new classes with limited labeled videos, ensuring these classes are distinct. Current methods [54, 53], typically use 2D-CNNs to extract features from individual frames and apply metric learning with temporal alignment to classify videos based on their similarity. While effective for centralized, single-machine learning, these methods struggle with capturing the temporal dynamics between frames and aren't designed for distributed environments. In response, our focus is on developing a Federated Learning (FL) framework for meta-training spatio-temporal models across distributed data sources, termed Federated Few-Shot Learning [10]. This approach aims to enhance FL's feasibility by ensuring privacy, reducing communication overhead, and addressing the challenge of domain variance across different clients' tasks.

This thesis introduces a novel framework, the Federated Learning for Action Recognition, which innovates in the realm of HAR by integrating both Supervised Learning and Few-shot Learning with FL. The framework initiates by deploying either a Few-Shot Learning (FSL) or a supervised learning algorithm, to train a local model using each client's local videos. Regardless of the approach being FSL or supervised, these local models are subsequently aggregated on a federated server, which serves to refine and enhance the global model. Addressing the limitations of 2D-CNN backbones, our framework utilizes spatiotemporal deep networks, such as 3D-CNNs and Transformers, to achieve effective feature embedding and explore temporal dy-

namics between frames, facilitating robust video representation across diverse client domains. The balanced focus on both supervised and few-shot learning approaches ensures the applicability of this thesis study across a wider range of learning tasks and data types, highlighting its versatility and potential for broader impact.

In our research, we introduce the use of the benchmark HAR datasets within the context of federated learning, a domain where these datasets have not been previously explored. By integrating these well-established datasets into federated learning scenarios, we aim to set a benchmark for future studies. This novel application not only extends the utility of the datasets but also provides a foundational framework for assessing the effectiveness and efficiency of federated learning algorithms in handling video and action recognition tasks. We believe this pioneering effort would be beneficial for future research, enabling a deeper understanding and further development of federated learning methodologies across diverse and complex datasets.

Our contributions are following:

- We propose benchmarking framework for efficient and privacy-preserving both supervised and few-shot action recognition.
- We conduct a comprehensive evaluation of various 3D-CNN, and Transformer based architectures as feature backbones, with a special focus on the effects of pre-training.
- We compare FL algorithms, specifically FedAvg and FedProx, in different realistic settings to optimize the performance.
- Our study demonstrates that the integration of pre-trained Transformer and 3D-CNN models with appropriate federated learning configurations can attain leading-edge results on challenging benchmark datasets for few-shot action recognition.

Our findings reveal that integrating pre-trained 3D-CNNs with optimal Federated Learning (FL) configurations not only enables our models to compete with their centralized learning counterparts in supervised learning scenarios but also, in some

instances, surpasses them in Few-Shot Learning (FSL) tasks. This improvement is notably significant given the inherent advantages of non-centralized data management, including enhanced privacy and data security. The adaptability of our FL framework to efficiently utilize distributed datasets showcases its potential to not just match but exceed the performance benchmarks set by centralized learning systems, especially in contexts where the robustness and generalization of models are crucial. This breakthrough sets a promising direction for future investigations in FL, FSL, and HAR, highlighting the viability of decentralized approaches in achieving superior performance across a spectrum of learning tasks.

Chapter 2

Background

2.1 Federated Learning

Federated Learning is a machine learning approach that enables the training of algorithms across multiple decentralized devices or servers holding local data samples, without the need to exchange or centralize these data samples. This approach contrasts with traditional centralized machine learning techniques, where all data is collected and processed on a central server. Instead, in Federated Learning, the model is sent to the device, where it learns from the data present, and only the model updates (and not the data itself) are sent back to a central server (Figure 2-1). These updates are then aggregated to improve the model, and the improved model is sent back to the devices. This cycle continues until the model achieves the desired level of accuracy. This method addresses privacy concerns, data security, and access rights, which are critical in sensitive applications such as healthcare, finance, and personal data analysis. The concept of FL was first introduced by McMahan et al. [28], who demonstrated its potential in training deep learning models on decentralized data efficiently and securely. Federated Learning systems are characterized by their ability to leverage diverse and rich data sources without compromising user privacy. They offer several advantages over traditional centralized learning frameworks, including reduced data transmission costs, enhanced privacy and security, and the ability to train models in real-time on edge devices. When applied to Human Action Recognition, FL

enables the utilization of data from diverse sources, such as smartphones, wearable devices, and surveillance cameras, without compromising user privacy. Federated learning was initially conceptualized to train deep learning models on decentralized data, with the aim of improving privacy and efficiency in model training. The approach aggregates locally computed updates (e.g., model gradients) rather than raw data, thus ensuring data privacy and reducing communication overhead [28].

2.1.1 Optimization Techniques

Federated learning involves challenges such as non-IID (independently and identically distributed) data, communication efficiency, and scalability. Research has focused on optimization algorithms like Federated Averaging (FedAvg) [28] to address these issues by efficiently aggregating model updates across distributed networks. Federated learning has evolved significantly with the development of various aggregation algorithms tailored to address the challenges of decentralized data training while ensuring privacy and efficiency. Federated Averaging (FedAvg), laid the groundwork for FL by aggregating model updates from distributed devices, proving effective for reducing communication costs despite struggles with non-IID data distributions. To enhance FL's adaptability to heterogeneous data, Federated Gradient and Knowledge Transfer (FedGKT) was proposed by in 2020 [17], focusing on a nuanced aggregation process that includes gradients and model outputs for improved generalization. Meanwhile, FedProx offers a solution to the non-IID challenge by incorporating a proximal term to maintain local models' proximity to the global model [25]. These algorithms collectively represent the forefront of FL research, addressing its core challenges and pushing the boundaries of privacy-preserving, efficient machine learning across decentralized networks.

Applying FL to action recognition involves training models to identify and classify actions in videos or real-time streams across multiple devices, such as smartphones, cameras, and IoT devices, without sharing the raw data. This application is particularly relevant for privacy-sensitive environments like smart homes or healthcare monitoring systems. Recent studies have showcased the effectiveness of FL in action

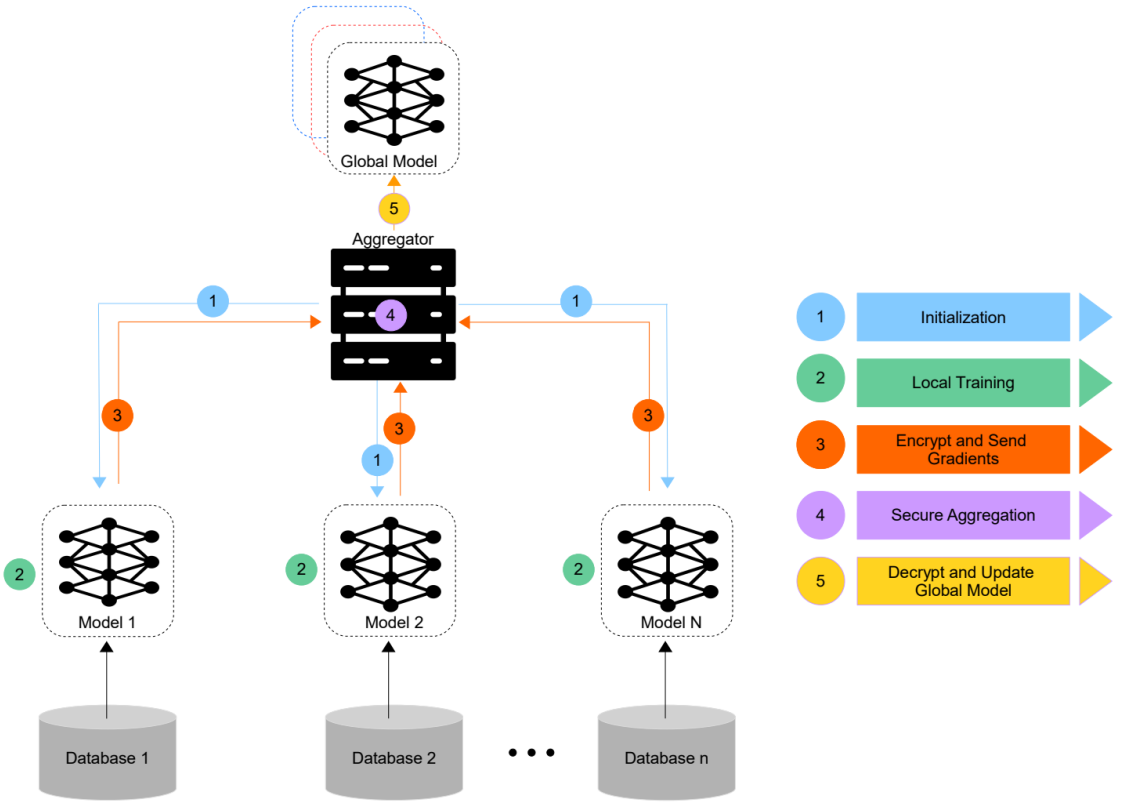


Figure 2-1: Federated Learning pipeline [19]

recognition, demonstrating its potential to facilitate real-time, privacy-preserving analytics across distributed networks [7, 52].

Federated Learning represents a significant shift in how data is utilized for machine learning, offering a pathway to harness the power of distributed data while upholding the principles of privacy and security. As FL continues to evolve, its application in fields like action recognition promises to unlock new possibilities for intelligent systems capable of learning from diverse, decentralized data sources without compromising user confidentiality.

2.2 Supervised Learning

Supervised Learning involves training models on a labeled dataset where each data sample is tagged with the action it represents. This method relies on large, annotated datasets to teach models the correlation between input features and output

labels (classes). The advantage of supervised learning is its ability to achieve high accuracy on well-defined tasks, given sufficient training data. The primary challenge in supervised learning is the dependency on extensive labeled datasets, which are costly and time-consuming to create. Despite this, the approach remains the gold standard for training models accross wide range of domains including HAR due to its effectiveness in learning complex patterns and nuances of the datasets.

2.3 Few-Shot Learning

Few-shot learning addresses the limitations of supervised learning by reducing the dependency on large annotated datasets. Few-shot learning aims to recognize actions from a minimal number of examples, leveraging prior knowledge and generalization capabilities of models. This approach is especially relevant in scenarios where collecting extensive labeled data is impractical or impossible. The essence of few-shot learning lies in its reliance on the model's ability to leverage prior knowledge and generalize from previous experiences to new, unseen tasks.

Few-shot learning can be further categorized into "N-shot" learning scenarios, where "N" refers to the number of examples per class that the model is exposed to during the training phase. For instance, in a "1-shot" learning scenario, the model learns from only one example per class. Similarly, in a "5-shot" learning scenario, the model has access to five examples per class. The challenge in N-shot learning is to design models and algorithms capable of extracting as much information as possible from these N examples to make accurate predictions on new, unseen data. Another important concept in few-shot learning is "N-way" learning, which refers to the number of classes involved in a given task. In an "N-way" classification task, the model is asked to classify input examples into one of N classes. Combining this with the concept of N-shot learning, we might have a "5-way 1-shot" learning task, where the model needs to classify examples into one of five classes, having seen only one example from each class during training. A key approach to addressing the few-shot learning challenge is meta-learning, or "learning to learn" [41]. Meta-learning

techniques involve training a model on a variety of learning tasks, such that it can learn new tasks quickly with only a few examples. The model effectively learns a learning strategy, allowing it to apply its prior knowledge to new problems efficiently. This can involve learning an optimal initialization of model weights, learning to adjust its learning rate, or learning how to select or construct features from new data in a way that facilitates quick learning.

Chapter 3

Related works

3.1 Human Action Recognition

Expanding on the initial overview in the earlier sections, this more detailed and comprehensive exploration of human action recognition includes recent advancements, methodologies, and key studies that have shaped the field. HAR is a dynamic research area in Computer Vision, with significant contributions across RGB-based, skeleton-based, and multimodal approaches. Recent trends emphasize deep learning, multimodal fusion, and the application of attention mechanisms for improved accuracy and efficiency.

The deep learning revolution has transformed HAR by introducing models that excel in capturing spatial and temporal features from video data. Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks for temporal dynamics, have been foundational [6, 21]. More recently, 3D CNNs and Transformer-based models have set new benchmarks for action recognition performance. Carreira and Zisserman introduced the Inflated 3D ConvNet (I3D) model, which extends 2D CNNs to 3D, significantly improving upon previous HAR methods by leveraging pre-trained ImageNet weights for video action recognition [5].

Transformers [40], originally designed for natural language processing tasks, have also been adapted for video understanding, offering an alternative to convolutional

methods by focusing on self-attention mechanisms to capture long-range dependencies. Arnab et al. presented a comprehensive study on the effectiveness of Vision Transformers for action recognition, showcasing their potential to outperform CNN-based approaches on standard benchmarks [1].

Additionally, Graph Convolutional Networks (GCNs) have been explored for their ability to represent and analyze the skeletal movements of humans, providing a more structured approach to understanding human actions [49].

While Deep Learning has enabled significant progress, HAR faces ongoing challenges such as handling diverse and complex actions, background clutter, and camera motion. A cornerstone of supervised HAR is the availability of large-scale video datasets such as Kinetics, UCF101, and HMDB51, which have propelled advancements in the field [5], [36], [23]. The variability in human appearance, occlusions, and the need for real-time processing further complicate recognition tasks. Addressing these challenges, recent research has focused on few-shot learning, cross-domain adaptation, and efficient network architectures to improve robustness and efficiency.

3.1.1 Supervised HAR

HAR aims to analyze and interpret human behaviors from visual data, facilitating applications in various domains, including security, healthcare, and entertainment. The transition from handcrafted feature extraction to deep learning has led to substantial improvements in recognition performance.

RGB-Based HAR

RGB-based approaches utilize color video data, focusing on capturing the spatial and temporal patterns indicative of human actions.

3D ResNets extend the successful ResNet [18] architecture to video data by replacing 2D convolutional layers with 3D convolutions, allowing the network to learn spatiotemporal features directly from video clips [38]. This approach enables the model to capture motion information and temporal context inherent in video se-

quences, making it well-suited for HAR tasks. 3D ConvNet models have been widely adopted in the community for their simplicity and effectiveness across a variety of video understanding tasks. Carreira and Zisserman introduced Inflated 3D ConvNet (I3D) that inflates filters and pooling kernels of 2D CNNs into 3D, enabling the network to learn rich spatiotemporal features [5]. Introduced by Feichtenhofer et al., SlowFast networks are based on the principle of processing video data at two different temporal resolutions [13]. The "Slow" pathway captures spatial semantics at a low frame rate, while the "Fast" pathway captures motion at a high frame rate. This dual-pathway architecture allows the model to efficiently integrate detailed spatial information with dynamic temporal information, leading to significant improvements in action recognition tasks.

Vision Transformers (ViTs) have been adapted to video understanding tasks, including HAR, by processing sequences of image frames as a series of patches to capture long-range dependencies [8]. The Multiscale Vision Transformer (MViT) incorporates a hierarchical structure with an efficient attention mechanism to model interactions across different scales and resolutions [12]. MViT dynamically adjusts the resolution of feature maps across different layers, enabling the model to focus on relevant spatiotemporal features while being computationally efficient.

VTN (Video Transformer Network) enhances RGB-based human action recognition by applying transformer principles to both spatial and temporal video dimensions [29]. It treats video frames as sequential patches, enabling it to capture complex action sequences through long-range dependency modeling. VTN complements existing technologies like 3D ConvNets and SlowFast networks, providing a nuanced approach to video understanding that leverages the transformer's strengths in handling multi-dimensional data. This integration marks a significant step forward in refining the capabilities of action recognition systems.

Skeleton-Based HAR

Skeleton data, representing human figures as a series of interconnected joints, provides a high-level, compact representation that is robust to variations in appearance and

environment.

1-dimensional CNN based skeleton model Double-feature Double-motion Network (DD-Net) presents a compact and fast solution for skeleton-based action recognition, focusing on addressing the common issues of large model size and slow processing speeds [50]. Utilizing a minimalistic design with only 0.15 million parameters, DD-Net achieves remarkable processing speeds—up to 3,500 FPS on GPUs and 2,000 FPS on CPUs. It employs a Joint Collection Distances (JCD) feature for location-viewpoint invariant information and a two-scale global motion feature to adapt to motion scale variances, simplifying the network structure while maintaining high performance. Demonstrating state-of-the-art results on JHMDB dataset [9] by achieving 77.2% of accuracy, DD-Net proves its efficacy and potential for practical applications in various multimedia fields.

GCNs, another very popular approaching when processing skeleton data, exploit the natural graph structure of skeleton data, with significant advancements including the introduction of spatial-temporal graph convolutional networks (ST-GCN) to model the dynamics of human actions effectively [49].

Transformers also do play a significant role in contribution to skeleton data processing. Recent studies have applied attention mechanisms and Transformer models to skeleton-based HAR, focusing on the temporal dynamics and relationships between different body parts for improved recognition accuracy [32].

Multimodal HAR

Integrating data from multiple modalities, such as RGB, depth, and skeleton information, offers a holistic view of human actions, leading to more robust recognition systems. Research has explored various fusion strategies (early, late, and hybrid) to combine features from different modalities, aiming to leverage the complementary information [33]. Techniques like cross-modal distillation enable learning richer representations by transferring knowledge between modalities, enhancing performance even in scenarios where some modal data might be missing or noisy [15].

3.1.2 Few-Shot Action Recognition

In the realm of few-shot action recognition, researchers have innovated with metric-based meta-learning approaches that construct a generalized metric space for comparing videos of different actions, using 2D-CNNs for feature extraction. Once the features are extracted, the distance between the distance between the features of the query and support video samples is being calculated. Notably, the Compound Memory Network (CMN) [53], OTAM [4], TRX employing CrossTransformers [31], and the Hybrid Relation Score Module (HyRSM) by Wang et al. exemplify advancements in this field [43].

The abovementioned methods employ the image-level feature extractions, that may not effectively capture the whole range of crucial temporal information. The spatiotemporal models such as R(2+1)D [38] used in TSL [46] and C3D [37] used by both TARN [3] and CMOT [27] meta-learning techniques respectively, on the other hand, were developed exactly for enhanced video-level feature generation. Other studies, such as [24] and [14], have explored the expansion of data samples through methods like data augmentation or the generation of new samples using generative models. Specifically, [24] utilized a conditional GAN, ProtoGAN, to augment the number of samples for each class within the support set. Conversely, [14], through AMeFu-Net, leveraged a novel approach by integrating depth and visual information, enhanced by temporal asynchronization augmentation. Providing very promising benchmark results, however, the centralized data processing of these techniques raises significant privacy and communication challenges, underscoring the untapped potential of federated learning in few-shot action recognition.

Few-shot learning in HAR involves training models to understand new actions from only a few labeled instances, relying on techniques such as meta-learning or transfer learning. Vinyals et al. introduced the concept of matching networks, a meta-learning approach that has been influential in few-shot learning research, including its application in HAR [41].

A significant advantage of few-shot learning is its potential to make HAR systems

more adaptable and efficient in learning new actions with limited data. However, the challenge lies in developing models that can effectively transfer knowledge across different actions and contexts.

Semi-supervised and few-shot learning approaches in HAR are still under active research, with the goal of overcoming the limitations of fully supervised methods. These strategies open up new possibilities for HAR applications where acquiring large labeled datasets is challenging, such as in personalized healthcare monitoring or in scenarios requiring rapid adaptation to new actions.

3.2 Federated Action Recognition

Federated learning (FL) is revolutionizing privacy in computer vision through federated action recognition, leveraging FL and distributed learning to safeguard data privacy. This approach prioritizes data privacy by processing data on local devices or servers rather than centralizing it, making it especially valuable in surveillance, healthcare, and other privacy-sensitive areas. It enhances secure and efficient model development for recognizing actions. Doshi et al. in 2022 and Zhang et al. in 2021 explored FL for identifying distracted driving using FedAvg with 2D-CNNs, with the former also incorporating FedGKT to optimize resource use on edge devices [7, 52]. Xiao et al. (2021) conducted a study on Federated Learning for recognizing human actions through wearable sensors [47]. They created a unique method that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) with an attention mechanism, alongside FedAvg and homomorphic encryption. Their approach surpassed previous human action recognition algorithms across various datasets. In another study provided by Ouyang et al. in 2022, the authors explored the use of Federated Learning (FL) for recognizing human actions [30]. They introduced a system named ClusterFL, which improves upon traditional FL by leveraging the natural grouping of user data. This clustering approach helps in boosting both the accuracy of the model and the efficiency of communication between devices.

Chapter 4

Methodology

This chapter describes the methodology employed in the experiments, with a particular focus on the Deep Learning models deployed, the utilization of open-source datasets, the configuration of federated learning and few-shot learning scenarios, and detailed training specifications. Through this comprehensive approach, we aim to offer a clear and thorough understanding of the experimental framework, ensuring that the insights gained are both robust and replicable.

4.1 General Framework

Our proposed method, is outlined in a step-by-step guide referred to as Algorithm 1. It introduces a specialized Federated Learning (FL) technique designed for collaborative learning across multiple clients. It works through a series of communication rounds, during which it coordinates the simultaneous enhancement of models by each participant. Initially, each client selects specific learning scenarios from their own data collection. Then, they proceed to refine their models by engaging in a learning process. In the Federated Learning framework, the individual models from each client, denoted as $\{\Theta^{(k)}\}$, are sent over to the central federated server. Here, they're combined to shape the overarching global model, $\{\Theta\}$. This unified model is then shared back with all the clients for use in future rounds, ensuring that each participant benefits from the collective learning. When it comes to applying this collective

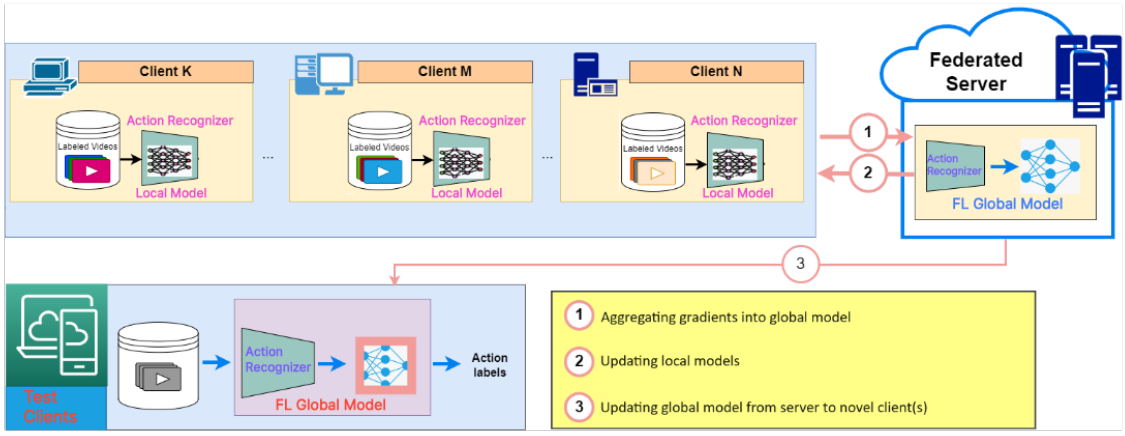


Figure 4-1: Federated Learning Framework

knowledge, the global model Θ is adapted to tackle and fine-tune new samples presented by novel clients, showcasing the model's ability to evolve and apply its learning to fine-tune new tasks.

4.1.1 Model Aggregation Algorithms

Aggregation algorithm that specifies the nature of aggregation of multiple locally trained models into a single global model is a crucial part in a federated learning system. For this Thesis, we have utilized two popular algorithms, namely FedAvg and FedProx.

FedAvg [28]. As already have been described in the Background section, the FL process starts by partitioning a all the data samples among client participants, where each participant updates the local model on its corresponding data portion in order to minimize the loss function derived locally. Federated Learning process is initiated by partitioning large dataset, and distributing it among local devices. The local models in the devices are being trained on the corresponding data portion and thus minimizing the local loss function. The updated models from all the participating clients are being sent to the central server where they are being aggregated into a global model. FedAvg uses weighted averaging of the gradients for the aggregation purposes.

Algorithm 1: General Algorithm [39]

Input : Communication rounds T, Number of clients K, Dataset $(\mathcal{X}_b, \mathcal{X}_n)$, Batch Size B, Local Epochs E, and Learning Rate η
Output : Global Model Θ
Initialize : Θ_0
for $t \leftarrow T$ **do**
  for *each client k in parallel* **do**
    $\Theta_t^{(k)} \leftarrow$ **ClientUpdate**(Θ_t)
  end
  Clients send model parameter $\{\Theta_t^{(k)}\}_{k=1}^{K}$ to server for aggregation:
  $\Theta_{t+1} = \sum_{k=1}^{K} \frac{|\mathcal{X}_b^{(k)}|}{|\mathcal{X}_b|} \Theta_t^{(k)}$
end
Return Θ_t
ClientUpdate():
  Input: global model from previous round Θ_t
  Output: updated local model $\Theta_t^{(k)}$
    $\mathcal{B}_k = \{\mathcal{T}_1, .., \mathcal{T}_m\}$
  Optimize $\Theta_t^{(k)}$ using local training process
  Return $\Theta_t^{(k)}$

FedProx [25] extends the Federated Averaging (FedAvg) algorithm to address challenges in federated learning, particularly when dealing with non-IID data across clients . The core idea behind FedProx is to introduce a proximity term to the local objective functions of the clients. This proximity term penalizes the divergence of the local models from the current global model. The FedProx update rule can be represented as follows:

$$F_i(w) = F_i(w; w^t) + \frac{\mu}{2}|w - w^t|^2 \tag{4.1}$$

where $F_i(w)$ is the objective function for the i-th client, which includes both the empirical loss on the client's local data and a proximal regularization term. w represents the parameters of the local model on the i-th client. w^t denotes the global model parameters at iteration t, which are shared with the clients at the beginning of the training round. $F_i(w; w^t)$ is the local empirical loss function on the i-th client. μ is a hyperparameter that controls the strength of the proximal term. $\|w - w^t\|^2$ is the squared Euclidean norm, measuring the distance between the local model parameters

w and the global model parameters w^t. This modification aims to make the algorithm more robust to the heterogeneity in the data and the systems of the clients.

4.2 Supervised Learning HAR with FL

The Supervised Human Action Recognition (HAR) pipeline utilizes raw video clips, which are preprocessed by sampling, resizing, cropping, and rescaling. The processed videos can either be directly analyzed using spatio-temporal models to capture RGB data dynamics, or converted into skeletal data for detailed structural analysis. In our study, we employ both approaches: direct processing of RGB data and extraction of skeleton data, feeding them into Spatio-Temporal neural networks to classify human actions effectively based on movement patterns and visual features. Utilized in the Federated Learning framework, this pipeline is depicted as Action Recognizer in the Figure 4-1.

4.2.1 RGB-Based HAR

In our study, we explore the application of advanced deep learning models for Human Activity Recognition (HAR) across two significant datasets, KTH [34] and JHMDB [9], leveraging both traditional centralized learning approaches and federated learning (FL) settings. We employ the R3D-18 (3D ResNet) model (Figure 4-2) pre-trained on the Kinetics-400 dataset, renowned for its efficacy in capturing spatial and temporal features from video data, to analyze and classify activities within the KTH and JHMDB datasets under a conventional centralized learning framework [38]. This initial phase serves as a baseline to assess the model's performance in recognizing a diverse array of human actions from video inputs without the complexities introduced by FL scenarios. Subsequently, we transition to a federated learning paradigm to train the R3D model across distributed clients. This approach allows us to evaluate the model's adaptability and performance under both IID and non-IID data distributions, reflecting real-world scenarios where data heterogeneity and distribution imbalances are prevalent. Such FL settings aim to enhance privacy and data security while assessing

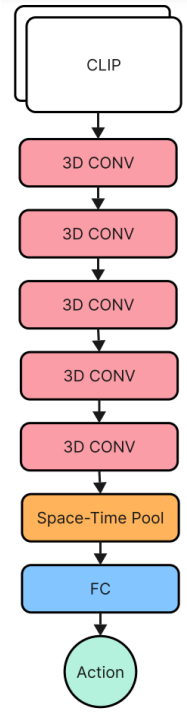


Figure 4-2: R3D Architecture [38]

the impact of data distribution on model accuracy and generalizability.

4.2.2 Skeleton-Based HAR

Further advancing our methodology, we employ DD-Net, a Double-feature Double-motion Network, designed specifically for skeleton-based activity recognition [50]. This 2-stream model utilizes a Joint Collection Distances (JCD) feature to achieve location-viewpoint invariant information, alongside a dual-scale global motion feature designed to accommodate variations in motion scale. Additionally, we have employed Transformer-based architecture. The architecture is structured around a Transformer encoder, consisting of a single encoder layer that employs a multi-head attention mechanism with four heads, allowing the model to concurrently focus on various parts of the input sequence for a nuanced understanding [2]. These 2 models are applied to the same KTH and JHMDB datasets to capture and classify activities based on skeletal movement data, thereby providing a complementary perspective to the R3D-18 model's video-based analysis.

4.2.3 Multimodal HAR

Finally, to leverage the strengths of both R3D and DD-Net models and improve the overall accuracy and robustness of activity recognition, we propose a fusion strategy for multimodal classification. For this Thesis, both early and late fusion strategies have been utilized. The late fusion approach integrates the predictions from both models by calculating the mean values of the prediction vectors derived from both models. In the early fusion approach, feature vectors from the models are concatenated prior to the fully connected layers of the models. These two fusion methods allow the learning system to take into account the rich spatial-temporal information extracted from video frames by the R3D model and the precise skeletal movement patterns captured by the DD-Net. By combining these two diverse but complementary sources of information, our multimodal classification framework aims to achieve superior performance in HAR tasks, effectively addressing the challenges posed by the KTH and JHMDB datasets under various learning settings. This comprehensive methodology not only tests the limits of individual models in isolated and federated learning scenarios but also explores the synergistic potential of combining different data modalities for enhanced activity recognition.

4.3 Few-shot Learning HAR with FL

In the realm of Federated Few-Shot Learning, the goal is to master new categories with only a handful of examples across various clients, denoted as $\{\mathcal{C}_k\}_{k=1}^{K}$, alongside a central federated server $\mathbb{S}$. Inside each client $\mathcal{C}_k$, Few-Shot Learning segregates the classes into two distinct groups: the base classes ($\mathcal{X}_b^{(k)}$), which have plenty of labeled examples for training, and the novel classes ($\mathcal{X}_n^{(k)}$), which are new to the model and come with only a few samples. This setup paves the way for a meta-learning strategy that revolves around episodic training. Referring to the Figure 4-1 in the context few-shot learning FL setup, the Meta-Learning process is denoted as a Action Recognizer. Here, each episode crafted within a client $\mathcal{C}_k$ pulls from the $\mathcal{X}_b^{(k)}$ pool.

Each episode, tailored to a specific learning task, is comprised of a support set

$S^{(k)} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M \times P}$ and a query set $Q^{(k)} = \{\mathbf{x}_j\}_{j=1}^{L}$. The support set is populated with labeled examples spanning M different classes, each represented by P samples. Meanwhile, the query set includes L unlabeled samples that are up for classification in the episode, setting up an M-way P-shot scenario.

Within an episode for client $\mathcal{C}_k$, the aim is to fine-tune a model $\Theta^{(k)}$ to accurately classify the samples in the query set $Q^{(k)}$ using insights gleaned from the support set $S^{(k)}$. This involves minimizing the discrepancy between the predicted and actual labels of $Q^{(k)}$. In the Federated Learning (FL) framework, these local models $\Theta^{(k)}$ are compiled on the federated server to construct an aggregated global model Θ, which is then disseminated back to the clients for future rounds.

When it comes time to test, the globally harmonized model Θ is adapted to take on new challenges presented by novel clients, showcasing its ability to apply learned knowledge to fresh situations.

In our approach, we harness the power of Prototypical Networks (ProtoNet) to navigate the meta-learning landscape across various client bases [35]. ProtoNet, at its core, focuses on crafting a prototype, essentially the average of feature vectors, for each base class. This model, serving as a metric-based classifier, is adept at gauging the similarity between examples in the support set and those in the query set throughout the meta-learning journey. For the sake of clarity, let's zoom in on the process within a single client, given that the meta-learning algorithm unfolds in a uniform manner across all clients in the federation. ProtoNet embarks on its task with an episode that includes a support set $\mathcal{S}$ and an unlabeled query sample $\mathbf{q}$, executing the following pivotal steps.

The initial step involves determining the prototype representation $\mathbf{p}_c$ for each class c within the episode. This is achieved by calculating the mean values of the feature representations of the support samples that belong to that class:

$$\begin{aligned} \mathbf{p}_c &= \frac{1}{N_c} \sum_{(\mathbf{x},y) \in \mathcal{S}_c} f(\mathbf{x}) \\ \mathcal{S}_c &= \{(\mathbf{x}, y) \in \mathcal{S} : y = c\} \end{aligned} \tag{4.2}$$

Here, $\mathcal{S}_c$ denotes the collection of support samples for class c, with $\mathbf{x}$ as a sample, y representing its class label, and $f(\cdot)$ being the feature extraction model.

Moving forward, the similarity between $\mathbf{q}$and each $\mathbf{p}_c$ is assessed using a distance metric, such as the Euclidean distance:

$$d(\mathbf{q}, \mathbf{p}_c) = \|f(\mathbf{q}) - \mathbf{p}_c\| \tag{4.3}$$

The final stage leverages the softmax function across the negative distances between the query sample and the prototypes to calculate the probability of the query sample $\mathbf{q}$ that is relevant to class c.

$$P(y = c|f(\mathbf{q}), \mathcal{S}) = \frac{\exp(-d(\mathbf{q}, \mathbf{p}_c))}{\sum_{c'} \exp(-d(\mathbf{q}, \mathbf{p}_{c'}))} \tag{4.4}$$

Through these steps, ProtoNet formulates class probabilities for each query sample, which are instrumental in training the feature extractor f among the base classes during meta-training, or assessing the model's performance on novel classes during meta-testing.

4.3.1 Spatio-Temporal Feature Backbones

For the experiments focusing on Few-shot learning, our methodology integrates four advanced 3D-CNN architectures to serve as the primary feature extractors. These models include:

- **R3D-18** [38]: A Residual 3D Network that utilizes 3D convolutions to capture spatial and temporal information in video data. It is a variant of the ResNet architecture adapted for video processing, making it efficient for action recognition tasks.

- **I3D** [5]: This model inflates filters and pooling kernels of pre-trained 2D CNNs (like those from ImageNet) into 3D, allowing for the learning of spatiotemporal features by leveraging the successful architectures of 2D CNNs.

- **SlowFast** [13]: A novel approach that involves two pathways; the Slow pathway capturing spatial semantics and the Fast pathway that is used to capture motion at a finer temporal resolution. We specifically utilize the Slow pathway configuration, processing 8 frames per input, to emphasize spatial feature extraction.

- **R(2+1)D** [38]: This architecture splits the 3D convolution operation into spatial (2D) and temporal (1D) components, facilitating more effective learning of spatial and temporal features by reducing the model complexity.

In addition to these, we extend our experimentation to include transformer-based models, recognizing their growing prominence in capturing long-range dependencies:

- **VTN** (Video Transformer Network) [29]: Pre-trained on ImageNet, this model applies transformer architectures to video understanding, leveraging the power of self-attention mechanisms to process sequential video frames.

- **MViT** (Multiscale Vision Transformers) [12]: Pre-trained on the Kinetics-400 dataset, MViT employs a pyramid-like structure to efficiently handle multiple scales, capturing rich spatiotemporal features across different resolutions.

The implementation of these models and their pre-trained parameters were sourced from Pytorchvideo and Torchvision platforms, ensuring a robust experimental framework [11]. This choice of 3D-CNNs and transformer-based models, especially those pre-trained on video datasets like Kinetics-400 [5], marks a significant departure from the more traditional 2D-CNN feature backbones predominantly pre-trained on ImageNet for few-shot action recognition [26]. This strategic selection aims to harness the intrinsic advantage of 3D-CNNs and transformers in capturing the dynamic spatiotemporal patterns inherent in video data, setting a solid foundation for our exploration into few-shot learning within the domain of action recognition.

Chapter 5

Experimental setup

In this chapter, we delve into the experimental setup that underpins our investigation, focusing on the meticulous implementation details and the settings of our experiments. A critical aspect of this exploration involves a comprehensive overview of the datasets utilized, each chosen to highlight different facets of our study's objectives. Through a detailed presentation of the models' configurations, the training procedures, and the evaluation metrics, this chapter aims to provide a clear blueprint of the experimental architecture.

5.1 Experimental settings

5.1.1 Federated Learning Implementation Details

Our experiments spanned both IID (where data is uniformly distributed across clients) and non-IID datasets (where data distribution varies significantly among clients), providing a well-rounded analysis of how these models perform under different data conditions. We conducted our experiments with the number of clients set to 4 for training purposes for both supervised learning and few-shot learning experiments. Using 4 clients optimally balances computational diversity with manageability, providing a practical setup that simulates real-world scenarios without overwhelming the computational resources available. This number allows for exposure to a varied data

set, essential for developing robust models that perform well across different distributions and ensuring the statistical significance of the results. Importantly, given the limited number of classes in the training sets of all datasets, it's crucial to keep the number of clients not too high, ensuring that each client receives a sufficient variety of classes and samples. This is particularly important in few-shot learning experiments, where the minimum number of available classes should be five or more due to the 5-way setup used. Thus, setting the client count to 4 allows for a realistic distribution of classes and samples per client, matching real-world conditions effectively. However, for the Slow model in the few-shot learning setup, we also tested with training client numbers ranging from 2 to 32 to evaluate the impact of the number of training clients on the performance. In all of our federated learning experiments, we chose to evaluate the global model using exactly one client due to the uniformity of the testing data available to all clients. This decision was made to align our evaluation practices with those of other studies that may not be using federated learning approaches but are utilizing the same global test dataset. Additionally, given that all clients in our setup share identical hardware specifications, choosing one client for evaluation eliminates any potential discrepancies that might arise from hardware differences. Specifically for FedProx, we incorporated a proximal regularization term, setting the proximal μ parameter to 0.001. This value was chosen empirically, as it showed the best results in preliminary experiments using FedProx, allowing us to effectively explore the impact of regularization on the algorithm's performance.

Regarding data distribution across multiple devices, we meticulously partitioned our datasets to achieve IID by ensuring an even spread of all data classes across every client. This setup guarantees that each client receives a representative slice of the whole dataset. On the flip side, for non-IID configurations, our goal was to foster diversity and uniqueness by minimizing the overlap of shared classes between clients. This approach not only challenges the models to adapt to varied and distinct data scenarios but also mirrors the complexity and heterogeneity of real-world data distributions, pushing the boundaries of what our models can learn and how they can be applied in diverse settings.

5.1.2 Supervised Learning Implementation Details

For the tasks involving raw RGB data, the R3D-18 model was trained for a total of 40 epochs when the centralizing approach was used. The video clips have been sampled into 16 frames. The frames have been resized to the size of 256×256 and cropped to 224×224 pixels. For FL experiments, the models were trained for a total of 15 round, each corresponding to 3 epochs of local training. The accuracy metric was used for model performance evaluation.

Regarding the skeleton data, the models are also trained and evaluated under the federated learning configurations (FedAvg and FedProx, across IID and non-IID settings) to determine its performance in a distributed learning environment and to compare its effectiveness against the R3D model in recognizing human activities from skeletal data. The DD-net has been trained for a total of 600 epochs in the centralized learning settings, while for the FL experiments we have dedicated 150 rounds each having 4 epochs of local training. The transformer model has been trained for 40 epochs in the traditional learning, and for the FL training, 15 rounds by 3 epochs were used.

5.1.3 Few-Shot Learning Implementation Details

For the FSL training in the Federated Learning settings, we followed the video data augmentation and preprocessing guidelines as detailed in the TSN study [42]. Depending on backbone model requirements, the videos have been sampled to 8 and 16 frames. In turn, frames have been resized to the size of 256×256 and cropped to 224×224 pixels. Spatial and temporal information from the videos have been obtained by feature extraction using the 3D-CNN or the Transformer based models excluding the final classification layer. The models in FL experiments have been meta-trained for 150 (5-shot) and 200 (1-shot) rounds each lasting 4 epochs, and as a comparison the models were trained for 1000 epochs in the traditional centralized learning setup. Each epochs comprised of 100 randomly selected episodes (meta-training tasks). The training periods, including the number of epochs and rounds in both few-shot and

supervised learning experiments, were strategically defined based on a blend of optimizing computational resources and achieving model convergence. Preliminary experiments played a crucial role in determining these parameters, enabling us to tailor the training duration to balance between computational efficiency and the effectiveness of the learning process. By specifying predefined training periods, we ensured that our results are comparable across different models and specifications. This approach was crucial for maintaining consistency and reliability in our evaluations. Adam optimizer has been used for the optimization purposes during training and the learning rate was set to 10^{-5}.

The performance of our model was evaluated using the 5-way 1-shot and 5-shot accuracy metric, which gauges the model's ability to accurately predict the correct category out of five possible categories, given a limited number of samples for each category (1 or 5). This evaluation was based on the average accuracy calculated over 10,000 episodes randomly selected from the meta-testing dataset. This number of testing episodes for evaluation was chosen as it offers a robust sample size that is large enough to provide statistically significant results. Additionally, using 10,000 episodes aligns with the standard practices in previous work within the field, allowing for direct comparability of our results with those of other studies [54].

In our research, we explored two primary model setups for Few-Shot Learning (FSL). Initially, we utilized models with random weights, meaning they were not pre-trained. This setup allowed us to assess the baseline learning capabilities of our models, starting entirely from scratch. The objective was to understand how effectively these models could learn without any pre-existing knowledge or data influence and evaluate the effect of. Additionally, we experimented with models pre-trained on the Kinetics-400 (K400) dataset. This second approach leveraged a rich pre-existing dataset to imbue our models with a broad base of knowledge, aiming to enhance their ability to quickly adapt to new tasks. The approach of using pre-trained models for this specific domain is aligned with relevant experimental methodologies provided in this particular research domain [54, 53]. The use of pre-trained models explored the impact of having a substantial foundational understanding from a large, diverse video

dataset on the efficiency of learning in FSL scenarios.

5.2 Datasets

5.2.1 Few-Shot Learning Datasets

For the Federated Learning experiments in the Few-shot learning settings, I have used 2 benchmark datasets to evaluate the performance. First is the Kinetics-100 dataset [53], which is a subset of Kinetics [5] dataset derived specifically for few-shot learning experiments. The dataset consists of video clips of total of 100 categories, split to the training, validation and testings sets as follows: 64 for meta-training, 12 for meta-validation and 24 for meta-testing. The Something-Something-V2 (SSV2) dataset is split similarly following the same distribution pattern [16].

Both SSV2 and Kinetics-100 are commonly used in few-shot learning research, particularly for action recognition [54, 53, 53, 4, 27]. This widespread usage makes them excellent benchmarks for assessing and comparing the performance of few-shot learning models, providing a solid foundation for validation against existing methodologies and results. SSV2 and Kinetics-100 encompass a wide range of action categories, which is critical for testing the versatility and robustness of few-shot learning algorithms. Kinetics-100, being a subset of the larger Kinetics dataset, offers a focused yet diverse set of actions, which complements the detailed and varied human-object interactions captured in SSV2. The popularity of these datasets in recent studies allows for benchmarking against a large body of research, facilitating comparisons that can drive innovations in model development. Using these well-recognized datasets ensures that the findings are relevant and contribute to the ongoing discussions in the research community about improving few-shot learning techniques. Some training samples of the aforementioned dataset are illustrated in the Figure 5-1.

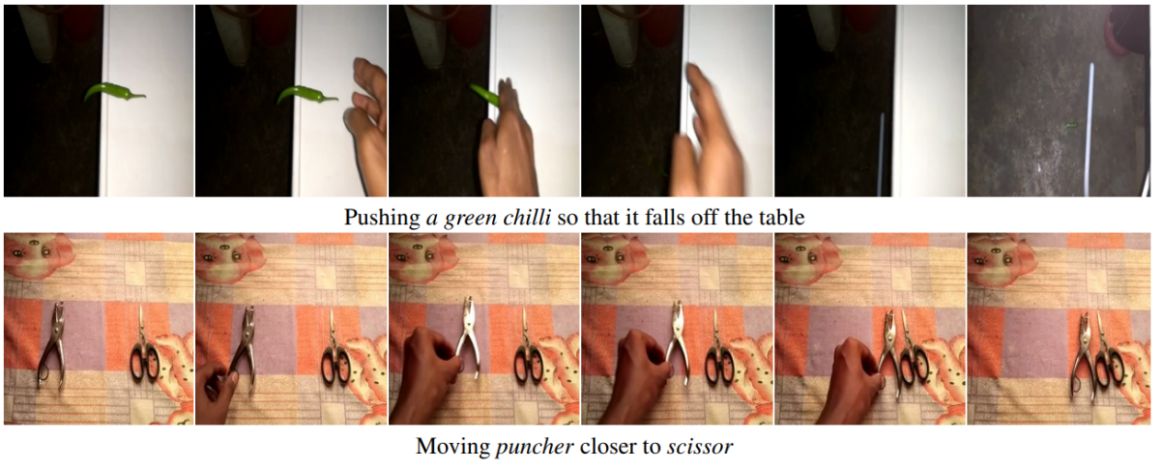


Figure 5-1: Something-Something-V2 frame samples [16]

5.2.2 Supervised Learning Datasets

The KTH dataset, one of the earliest benchmarks in HAR [34], offers 6 types of human activities performed by 25 subjects in different settings, providing foundational challenges in terms of variability and complexity. Videos in the dataset are averagely 10 seconds in length. Similarly, the JHMDB dataset, with its focus on human motion and its annotation of 21 distinct activities, enriches the pool of data with more nuanced, real-world actions [9].

Both the KTH and JHMDB datasets are known for their diversity in terms of actions and scenarios, as well as the quality of video data. KTH includes basic human movements which help in establishing a strong foundational model, while JHMDB contains more complex and natural human activities that are challenging and closer to real-life scenarios. Using both datasets allows for training models on a broad spectrum of human activities—from simple movements in controlled environments (KTH) to complex interactions in varied settings (JHMDB). This diversity is crucial for developing a robust model that performs well across different visual contexts, making the data ideal for validating the effectiveness of models accepting different modality.

Chapter 6

Results

This chapter delves into the critical outcomes and insights garnered from the previously discussed experiments, highlighting the pivotal achievements of our research. It offers a detailed comparative analysis, drawing clear distinctions in accuracy between the newly proposed system and the pre-existing methodologies. Through this analysis, readers will gain a comprehensive understanding of how our approach stands out in terms of performance and efficiency in certain experiments, marking a significant advancement in the field.

6.1 Supervised Learning

The results presented below commence with the results for the experiments done in the supervised learning settings. The section compose the accuracy results for the video action classification for the RGB, Skeleton, and Fusion-based models.

6.1.1 RGB-based HAR

The table 6.1 details the accuracy results for R3D-18 3D-CNN model tested on the JHMDB and KTH datasets that composes of video clips of 21 and 6 classes respectively. Notably, the model has been initialized with the weights pre-trained on the expansive Kinetics-400 dataset. The model was evaluated under both centralized

Table 6.1: Accuracy Results for R3D-18 CNN model (4 clients)

Dataset	Centralized	IID	Non-IID
JHMDB	71.18%	70.44%	64.56%
KTH	86.11%	83.33%	78.70%

learning and Federated Learning (FL) setups with data distributed in IID (Independently and Identically Distributed) and Non-IID formats across 4 clients using FedAvg aggregation method.

For the JHMDB dataset, the R3D-18 model achieved an accuracy of 71.18% in a centralized learning environment. When the data was distributed in an IID manner across clients, the model's accuracy slightly decreased to 70.44%. The decrease was more pronounced under Non-IID conditions, where the accuracy dropped to 64.56%, indicating challenges in learning from non-identically distributed data.

Similarly, on the KTH dataset, the centralized learning accuracy was 86.11%, which is notably high. However, under the IID distributed scenario, the accuracy fell to 83.33%, and further down to 78.70% in the Non-IID scenario.

These results underscore the impact of data distribution on the performance of FL models, particularly highlighting the challenge posed by Non-IID data distributions. Despite these challenges, the R3D-18 CNN model demonstrates robust performance across different setups, making it a viable option for federated human action recognition tasks.

6.1.2 Skeleton-Based and Fusion HAR

The experiments conducted on the JHMDB and KTH datasets with DD-Net and Transformer models reveal nuanced insights into the performance of Federated Learning (FL) optimization strategies, FedAvg and FedProx, under different data distributions. The tables 6.2 and 6.3 present accuracy results for different skeleton models across two datasets, JHMDB and KTH, evaluated under centralized, IID, and Non-IID federated learning settings, alongside a fusion model analysis.

For the JHMDB dataset (Table 6.2), the DD-Net model demonstrates superior performance in the centralized setting with a 77.16% accuracy, followed closely by its

IID configuration at 74.56%. However, a notable decline is observed under the Non-IID scenario, dropping to 61.49% with FedProx, suggesting challenges in handling data heterogeneity. The Transformer model shows a similar trend, though starting from a lower centralized accuracy of 73.12%, it experiences significant performance deterioration in Non-IID conditions, down to 53.46% with FedProx, highlighting the impact of data distribution on learning efficiency. Conversely, on the KTH dataset, the DD-Net again leads in the centralized framework at 85.65% accuracy, with relatively lesser performance degradation across federated learning setups (Table 6.3). Interestingly, the model maintains commendable resilience in Non-IID conditions, achieving 79.16% with FedProx, indicating better adaptability to data variability. The Transformer model, starting with a higher centralized accuracy of 86.57%, also shows robustness in federated settings, particularly with FedProx under IID and Non-IID conditions, reaching up to 84.43% and 77.57% respectively, which underscores the potential of advanced model architectures in distributed learning environments. The late fusion model (Table 6.4), combining R3D-18 and DD-Net on the KTH dataset, yields the highest centralized accuracy of 87.03%, with performance gracefully declining to 83.33% in IID settings and 78.97% under Non-IID. The fusion model also shows the best results for the JHMDB dataset achieving the 78.22% of accuracy in the centralized learning settings and 75.16% in the FL settings. This suggests that model fusion in a federated learning setting, while potentially resulting in lower accuracy compared to centralized learning methods, can still maintain robustness by leveraging the complementary strengths of individual models.

Key findings indicate that while centralized training generally offers the best performance, federated learning setups, especially with advanced algorithms like FedProx, present viable pathways to mitigating performance loss due to data distribution challenges. The resilience of the DD-Net and Transformer models, alongside the promising results from the fusion approach, underline the strategic importance of model choice and fusion strategies in optimizing federated learning outcomes.

The table 6.5 presented outlines the accuracy results for an early fusion model involving the models and datasets mentioned previously. From the data, it's evident

Table 6.2: Accuracy Results for Skeleton Models for JHMDB

Model	**Centralized**	**IID**		**Non-IID**	
		FedAvg	**FedProx**	**FedAvg**	**FedProx**
DD-Net	77.16%	74.56%	71.71%	64.41%	61.49%
Transformer	73.12%	64.85%	66.92%	51.12%	53.46%

Table 6.3: Accuracy Results for Skeleton Models for KTH

Model	**Centralized**	**IID**		**Non-IID**	
		FedAvg	**FedProx**	**FedAvg**	**FedProx**
DD-Net	85.65%	78.81%	80.18%	78.66%	79.16%
Transformer	86.57%	82.12%	84.43%	72.93%	77.57%

that the highest accuracy is achieved in the Centralized setting for both datasets, with a noticeable drop in performance in the IID setting, and the lowest accuracy observed in the Non-IID setting, similar to the late fusion model. The table indicates that the early fusion approach might be less effective than late fusion, where features from individual models are combined at a later stage in the processing pipeline.

6.2 Few-shot Learning HAR

For the Few-shot Learning experimental setup, we present our findings in tables 6.6 and 6.7 , focusing on the performance of 4 different 3D-CNN architectures: R3D, I3D, Slow, R(2+1)D, 2 Transformer based architectures: MViT and VTN. The models are tested under various data distribution conditions, including centralized, IID and non-IID scenarios. These experiments were conducted using varied number clients in federated learning (FL) setups. Table 6.6 explores the effectiveness of these models in both 1-shot and 5-shot learning tasks on the K100 dataset without any prior pre-training where 4 client participants have been used. Notably, Slow architecture generally outperforms the others, showcasing its superior capability in capturing detailed semantic information and motion variations within video frames. A noticeable trend across all models is a drop in accuracy when transitioning from a centralized to an FL setting, attributed to the diverse and heterogeneous nature of data across multiple clients, which can negatively impact FL performance in the K100 context. Furthermore, models tend to achieve better results in IID scenarios compared to non-

Table 6.4: Accuracy Results for the Late Fusion model

Model 1	**Model 2**	**Dataset**	**Centralized**	**IID**	**Non-IID**
R3D-18	DD-Net	KTH	87.03%	83.33%	78.97%
R3D-18	DD-Net	JHMDB	78.22%	75.16%	66.74%

Table 6.5: Accuracy Results for the Early Fusion model

Model 1	**Model 2**	**Dataset**	**Centralized**	**IID**	**Non-IID**
R3D-18	DD-Net	KTH	84.72%	81.02%	77.78%
R3D-18	DD-Net	JHMDB	75.59%	73.09%	65.21%

IID settings due to the latter's unbalanced class distribution. Particularly, in the IID scenario, R3D achieves an accuracy of 46.68% in the 5-shot task, while Slow stands out with 54.28% accuracy for the same task. Regarding the non-iid settings, VTN outperforms all the other feature extractors achieving 40.66% and 52.34% accuracy for 1-shot and 5-shot settings respectively. These outcomes highlight the distinct strengths of the VTN model across challenging data distribution scenarios, while Slow excelling in the less challenging IID experiments.

Table 6.6: Comparisons of different backbones on K100.

	Centralized		IID		non-IID	
Models	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
R3D*	46.16%	48.94%	41.33%	46.68%	36.24%	45.26%
I3D*	37.82%	47.78%	36.13%	44.44%	34.18%	42.12%
R(2+1)D	43.58%	57.74%	39.26%	53.46%	37.22%	52.12%
MViT	42.14%	58.24%	36.28%	52.12%	36.44%	52.00%
VTN	44.18%	57.64%	42.18%	53.90%	**40.66%**	**52.34%**
Slow*	**47.88%**	**59.16%**	**42.22%**	**54.28%**	36.46%	44.16%

* The results for models marked have been published for a workshop paper [39]

Table 6.7 delves into the performance of the same 3D-CNN and Transformer based models, now pre-trained, on the K400 dataset using 4 client for the FL experiments. Interestingly, with this setting, MViT attains a lead in performance in all of the scenarios. With pre-training, FL settings surprisingly exhibit higher accuracy than centralized settings, and performances in IID and non-IID scenarios are quite similar. These results affirm the benefits of FL for few-shot learning (FSL) tasks, particularly

Table 6.7: Comparisons of different backbones on SSv2.

	Centralized		IID		non-IID	
Models	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
R3D*	30.54%	33.48%	38.40%	49.08%	38.52%	49.78%
I3D*	41.12%	60.82%	43.36%	61.02%	42.26%	61.08%
R(2+1)D	45.00%	59.02%	45.22%	59.16%	45.98%	59.70%
MViT	**54.52%**	**70.04%**	**62.14%**	**77.34%**	**61.28%**	**77.82%**
VTN	44.94%	59.20%	43.12%	55.66%	43.24%	56.62%
Slow*	44.44%	60.66%	45.90%	60.82%	45.48%	61.74%
Slow(Random)*	23.28%	24.62%	32.90%	45.02%	35.88%	44.40%

* The results for models marked have been published for a workshop paper [39]

when leveraging pre-trained model backbones. Notably, the best performing model MViT achieved 77.82% of accuracy in 4 client Non-IID settings for 5-shot experiments compared to 70.04% reached during centralized training. Pre-training enables the transfer of semantic information and meta-knowledge across clients, enhancing the global model's ability to adapt to new tasks more efficiently. In essence, combining pre-training with FL significantly boosts the meta-learners' generalization capacity to novel clients and mitigates the issue of data heterogeneity among FL clients.

Further investigation into the impact of pre-training on few-shot action recognition using the SSv2 dataset reveals a stark contrast when employing the Slow model with random initial weights; accuracy drops drastically across all settings. Centralized accuracy approaches that of random guessing for 5-way tasks, aligning with outcomes reported in other meta-learning methodologies. This decline is largely due to SSv2's emphasis on complex spatiotemporal reasoning, making it difficult to learn discriminative features from scratch on this dataset. Remarkably, FL settings manage to achieve significantly higher accuracy than centralized approaches. The iterative aggregation of FL models facilitates the transfer of meta-knowledge among clients, aiding in more effective learning of spatiotemporal features.

The findings from our experiments, as detailed in tables 6.8 and 6.9 , compare the performance of the advanced FedProx algorithm to the FedAvg algorithm across the K100 and SSv2 datasets. For these experiments, we chose the Slow architecture as

Table 6.8: Comparisons of different FL methods on K100.

	IID		non-IID	
FL algs	1 shot	5 shot	1 shot	5 shot
FedAvg	**42.22%**	**54.28%**	**36.46%**	**44.16%**
FedProx	40.98%	51.00%	35.10%	43.18%

Table 6.9: Comparisons on different FL methods on SSv2.

	IID		non-IID	
FL algs	1 shot	5 shot	1 shot	5 shot
FedAvg	45.90%	60.82%	45.48%	61.74%
FedProx	**47.24%**	**65.68%**	**46.66%**	**65.68%**

our primary feature backbone and evaluated the models under both IID and non-IID data distributions among four clients. Our observations reveal that while FedProx did not outperform FedAvg on the K100 dataset, however, it significantly enhanced model performance on the SSv2 dataset. This improvement underscores the importance of pre-training not just for the effectiveness of 3D-CNN backbones but also for the efficiency of federated learning (FL) algorithms. Additionally, FedProx demonstrated a resilience to performance degradation in non-IID scenarios compared to IID scenarios, likely due to its proximal term, which helps mitigate the adverse effects of data heterogeneity among clients.

6.2.1 Comparison of Few-shot learning results with state-of-the-art

In this analysis, we benchmark our approach against a variety of leading methods in centralized few-shot action recognition. Our evaluation encompasses best configurations for our conducted experiments each harnessing the Protonet meta-learning algorithm using the Slow network and MViT underlying architecture. Our comparison includes baseline methods like Meta-Baseline and Baseline Plus, as well as cutting-edge techniques such as CMN, OTAM, CMOT, TRX, HyRSM, SloshNet, and TADRNet, with findings presented in Table 6.10.

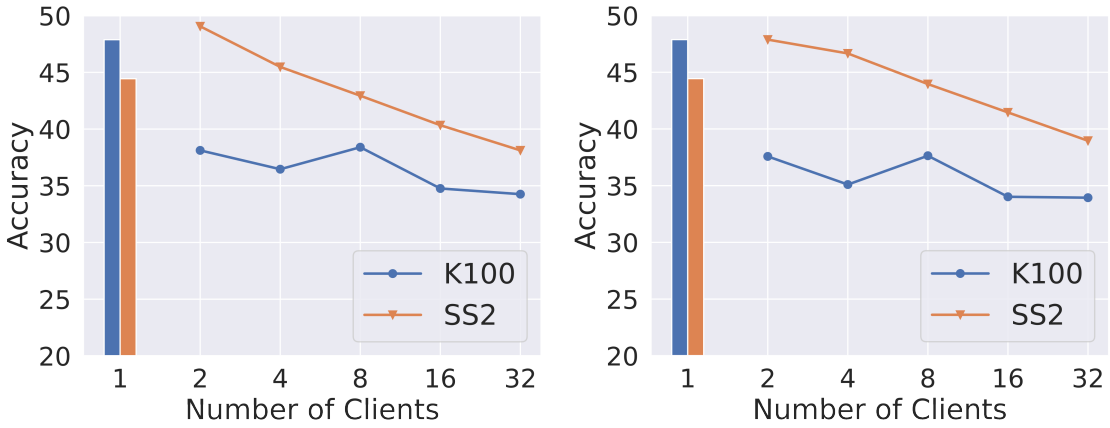
Table 6.10: Comparison with SOTA Few-Shot Action Recognition Methods.

Method	**Setting**	**K100**		**SSv2**	
		1 shot	5 shot	1 shot	5 shot
Meta-Baseline [54]	Centralized Learning	42.46%	49.78%	33.6%	43.0%
Baseline Plus [54]	Centralized Learning	<u>46.24</u>%	56.92%	46.04%	61.10%
CMN [53]	Centralized Learning	40.37%	50.27%	34.4%	43.8%
OTAM [4]	Centralized Learning	44.37%	50.07%	42.8%	52.3%
CMOT[27]	Centralized Learning	-	-	46.8%	55.9%
TRX [31]	Centralized Learning	-	-	42.0%	64.6%
HyRSM [43]	Centralized Learning	-	-	<u>54.3%</u>	<u>69.0%</u>
SloshNet [48]	Centralized Learning	-	-	46.5%	68.3%
TADRNet [44]	Centralized Learning	-	-	43%	61.1%
Proposed	Slow (best setting)	**47.88%**	**59.16%**	45.90%	61.74%
Proposed	MViT (best setting)	42.14%	<u>58.24%</u>	**62.14%**	**77.82%**

Note: Best and second-best results are denoted in bold and underlined, respectively.

For the K100 dataset, we align with the initial few-shot learning (FSL) scenario by not utilizing pre-training for any method, in line with the stance that using external pre-training data like K400 for backbone models contradicts the essence of FSL. Our method showcases superior performance in FSL settings when compared to others, with Centralized Learning using the Slow model emerging as the top performer within our variants. Notably, our approach, whether through FL or centralized learning, demonstrates enhanced effectiveness over Meta-Baseline, Baseline Plus, CMN, and OTAM. This is attributed to our use of 3D-CNN models, which excel in crafting more nuanced and discriminative representations for new tasks.

When it comes to the SSv2 dataset, we adopt a pre-trained MViT models for evaluation. Among our competitors, CMOT leverages a spatiotemporal network pre-trained on Sport-1M, whereas other top-tier methods opt for frame-level feature backbones pre-trained on ImageNet. Our findings indicate that framework presented using MViT model as a feature extractor with the IID setting for 1-shot and Non-IID setting using for the 5-shot learning using 4 clients for training surpasses all of the other mentioned methods. Intriguingly, certain FL configurations of our method even surpass the accuracy of leading centralized approaches, suggesting that FL's collaborative meta-knowledge aggregation across clients may enhance the model's generalization capabilities.



(a) FedAvg (b) FedProx

Figure 6-1: 1-shot learning accuracy results with clients from 1-32 (Slow model) [39].

6.2.2 Effect of different number of clients

To evaluate the impact of the number of clients used for training, we have exploited the Slow model for different FL experiments with number of devices ranging from 2 to 32. The experiments involved both datasets using FedAvg and FedProx algorithms under the non-IID setting. From the Figure 6-1, we can make a general claim that the larger number of clients lead to a decrease in the performance of the system. However, for the SSv2 dataset, as outlined earlier, FL experiments (utilizing 2 or 4 clients) can exceed the centralized learning in performance. The overall downwards trend, though, suggests that as data gets more fragmented across more clients, the models struggle to meta-learn effective global generalizations.

Chapter 7

Conclusion

The findings from our study indicate that the application of Federated Learning to Human Action Recognition, incorporating both supervised and Few-Shot Learning approaches, not only mitigates the limitations associated with centralized data processing but also enhances model performance and privacy preservation. By utilizing pre-trained 3D-CNNs and Transformer-based architectures within an FL framework, our models demonstrated robustness across various data distributions, achieving competitive performance against centralized learning systems. The results of this study position the Slow and MViT models used under federated learning settings as groundbreaking approaches for few-shot learning action recognition. These models not only align with but also surpass the current performance benchmarks, paving the way for significant advancements in federated learning technologies. Additionally, we propose that late fusion strategy combining RGB and skeleton data offers enhanced accuracy over individual models, presenting a compelling alternative for integrating multiple data sources or feature sets.

The integration of FL significantly contributes to the scalability, efficiency, and privacy of HAR systems, particularly in privacy-sensitive applications. Future research directions include exploring more advanced FL algorithms, further enhancing data privacy measures, and expanding the application of FL in HAR to include more diverse and complex scenarios. Our work sets a promising foundation for the advancement of FL in HAR, showcasing the potential of decentralized approaches in

achieving superior performance across a spectrum of learning tasks.

Looking ahead, future research will explore a broader spectrum of FL algorithms to identify and develop more sophisticated methods that can further enhance model performance, privacy, and improve model adaptability across diverse data landscapes. Additionally, employing a wider range of datasets, including those from emerging and underrepresented domains, will be crucial in testing the robustness and generalizability of FL models across diverse real-world scenarios. Another promising direction involves investigating various fusion methods to optimize the integration of different data types, thereby improving the accuracy and reliability of action recognition. These endeavors will not only contribute to the advancement of FL and HAR technologies but also pave the way for their application in increasingly complex and privacy-sensitive environments. By expanding the scope of FL algorithms, datasets, and fusion methods, future research will continue to push the boundaries of what is possible in HAR, opening up new opportunities for innovation and application in the field.

Bibliography

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[2] Aizada Askar, Min-Ho Lee, Thien Huynh-The, and Nguyen Anh Tu. 2d skeleton-based action recognition using action-snippets and sequential deep learning. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2372–2377. IEEE, 2022.

[3] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019.

[4] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *CVPR*, pages 10615–10624, 2020.

[5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[7] Keval Doshi and Yasin Yilmaz. Federated learning-based driver activity recognition for edge devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3338–3346, 2022.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Ujjal Kr Dutta, Mehrtash Harandi, and Chellu Chandra Sekhar. Unsupervised deep metric learning via orthogonality based probabilistic loss. *IEEE Transactions on Artificial Intelligence*, 1(1):74–84, 2020.

[10] Chenyou Fan and Jianwei Huang. Federated few-shot learning with adversarial learning. In *2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pages 1–8. IEEE, 2021.

[11] Haoqi Fan, Tullie Murrell, Heng Wang, Kalyan Vasudev Alwala, Yanghao Li, Yilei Li, Bo Xiong, Nikhila Ravi, Meng Li, Haichuan Yang, et al. Pytorchvideo: A deep learning library for video understanding. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3783–3786, 2021.

[12] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[14] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. Depth guided adaptive meta-fusion network for few-shot video recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1142–1151, 2020.

[15] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017.

[17] Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in Neural Information Processing Systems*, 33:14068–14080, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Yassine Himeur, Iraklis Varlamis, Hamza Kheddar, Abbes Amira, Shadi Atalla, Yashbir Singh, Faycal Bensaali, and Wathiq Mansoor. Federated learning for computer vision. *arXiv preprint arXiv:2308.13558*, 2023.

[20] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[21] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[22] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[23] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T." Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[24] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[25] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[26] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022.

[27] Su Lu, Han-Jia Ye, and De-Chuan Zhan. Few-shot action recognition with compromised metric via optimal transport. *arXiv preprint arXiv:2104.03737*, 2021.

[28] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[29] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021.

[30] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Guoliang Xing, and Jianwei Huang. Clusterfl: A clustering-based federated learning system for human activity recognition. *ACM Transactions on Sensor Networks*, 19(1):1–32, 2022.

[31] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *CVPR*, pages 475–484, 2021.

[32] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *Pattern recognition. ICPR international workshops and challenges: virtual event, January 10–15, 2021, Proceedings, Part III*, pages 694–701. Springer, 2021.

[33] Hossein Rahmani and Ajmal Mian. 3d action recognition from novel viewpoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1506–1515, 2016.

[34] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[35] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, page 4080–4090, 2017.

[36] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[38] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[39] Nguyen Anh Tu, Assanali Abu, Nartay Aikyn, Nursultan Makhanov, Min-Ho Lee, Khiem Le-Huy, and Kok-Seng Wong. Fedfslar: A federated learning framework for few-shot action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 270–279, January 2024.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Avukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, volume 29. Curran Associates, Inc., 2016.

[42] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018.

[43] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *CVPR*, pages 19916–19925, 2022.

[44] Xiao Wang, Weirong Ye, Zhongang Qi, Guangge Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Hanzi Wang. Task-aware dual-representation network for few-shot action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[45] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[46] Yongqin Xian, Bruno Korbar, Matthijs Douze, Lorenzo Torresani, Bernt Schiele, and Zeynep Akata. Generalized few-shot video classification with video retrieval and feature generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8949–8961, 2021.

[47] Zhiwen Xiao, Xin Xu, Huanlai Xing, Fuhong Song, Xinhan Wang, and Bowen Zhao. A federated learning system with enhanced feature extraction for human activity recognition. *Knowledge-Based Systems*, 229:107338, 2021.

[48] Jiazheng Xing, Mengmeng Wang, Yong Liu, and Boyu Mu. Revisiting the spatial and temporal modeling for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3001–3009, 2023.

[49] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[50] Fan Yang, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, pages 1–6, 2019.

[51] Guangle Yao, Tao Lei, and Jiandan Zhong. A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters*, 118:14–22, 2019.

[52] Bin Zhang, Jingya Wang, Junyi Fu, and Jinxiang Xia. Driver action recognition using federated learning. In *Proceedings of the 7th International Conference on Communication and Information Processing*, pages 74–77, 2021.

[53] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *ECCV*, September 2018.

[54] Zhenxi Zhu, Limin Wang, Sheng Guo, and Gangshan Wu. A closer look at few-shot video classification: A new baseline and benchmark. *arXiv preprint arXiv:2110.12358*, 2021.