Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ24664

Comparison of Vision Transformer and ResNet under Federated Learning in Human Action Recognition

Yicheng Li

School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai, China
221650335@stu.lixin.edu.cn

Abstract. This study investigates the adaptability of ResNet50 and Vision Transformer (ViT) in Federated Learning (FL) versus centralized training for human action recognition, addressing data privacy and distributed data silos in real-world applications. Employing the Kaggle dataset (about 12,000 images across 15 actions), the thesis standardized preprocessing (bilinear resizing, flip/rotation augmentation, ImageNet normalization) and conducted controlled experiments. Centralized training utilized full-batch gradient descent, while FL simulated non-independent and identically distributed (non-IID) data partitioning with Federated Averaging (FedAvg) aggregation. The results indicate that ViTs achieved 87.30% accuracy in FL, surpassing Residual Network (ResNet) by 8.01%. The train-validation gap was significantly reduced (1.67% vs. 11.35%), revealing improved generalization. Overfitting was limited through phased aggregation, which also enabled late-stage convergence exceeding. ViT's global attention proved more adaptable to heterogeneous data distributions, and FL's distributed optimization enhanced model resilience. These results highlight the complementary strengths of ViT and FL, especially for privacy-sensitive applications like healthcare, where preserving data sovereignty is essential.

Keywords: Federated Learning, Vision Transformer, Human Action Recognition, Data Privacy, Non-IID.

1. Introduction

Human Action Recognition (HAR) is a key task in computer vision, crucial for applications like intelligent surveillance, human-computer interaction, and remote health monitoring [1]. It aims to model human motion patterns from video or image sequences and accurately classify behaviors. A key challenge is to develop robust feature representations and efficient recognition frameworks that generalize well across diverse environments [2]. Recently, deep learning models such as Convolutional Neural Networks (CNNs) [3] and Vision Transformers (ViTs) [4] have shown impressive performance in HAR tasks under centralized training paradigms [5], especially when trained on large, annotated datasets. In real-world scenarios, however, HAR systems often face fragmented data across multiple devices or institutions, with collection constrained by privacy regulation, data sovereignty, and organizational policies. This results in data silos that impede practical model training and limit generalization performance. Federated Learning (FL) [6] is an

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.TJ24664

increasingly adopted approach that enables collaborative training by allowing individual clients to optimize local models and share only model updates with a central server. It will allow distributed clients to jointly train a global model without sharing raw data. By supporting decentralized optimization while preserving data privacy, FL is well-suited for privacy-sensitive HAR applications [7]. This study compares the performance of ResNet50 and ViT under federated and centralized training paradigms. A controlled experimental setup evaluates each architecture's response to data heterogeneity. The comparison focuses on communication efficiency and model convergence behavior over 200 training epochs. Results reveal architectural differences in adapting to distributed learning. These findings provide practical guidelines for model selection in edge-computing scenarios that require privacy-preserving learning.

In previous studies, Residual Network (ResNet) 50 has been widely adopted as a standard feature extractor in centralized training frameworks. Primarily attributed to its innovative residual connections that effectively mitigate the vanishing gradient challenge, thus ensuring robust performance in deep neural network architectures [8]. By preserving gradient flow throughout the network, residual connections have significantly enhanced model stability and training efficiency, thereby enabling the development of deeper and more sophisticated architectures. In contrast, ViTs, leveraging their global self-attention mechanisms, have emerged as formidable competitors in image classification tasks, demonstrating their potential to redefine the landscape of computer vision research [9]. Their ability to capture long-range dependencies and contextual information has led to state-of-the-art performance across various benchmarks. However, the integration and effectiveness of ViTs within distributed learning frameworks, particularly federated environments, remain a relatively uncharted territory. FL, with methodologies such as Federated Averaging (FedAvg), offers solutions for addressing non-Independent and Identically Distributed (non-IID) data distributions through parameter averaging across clients [10]. Despite these advancements, there is a conspicuous lack of comparative analyses between CNNs and Transformers under diverse data partitioning strategies, such as horizontal (sample-based) and vertical (feature-based) splits. Recent studies have endeavored to bridge this gap by introducing adaptive aggregation techniques tailored to ViT's attention heads to enhance their compatibility with federated optimization processes [11]. Nonetheless, these efforts often lack standardized benchmarks and rigorous empirical validation. Furthermore, the disparities in convergence dynamics, generalization capabilities, and gradient stability across model architectures and training methodologies underscore the necessity for systematic investigations.

This study develops an analytical framework linking model architectures to training approaches. Standardized Kaggle dataset experiments apply bicubic interpolation for resolution consistency. Data augmentation employs random flipping and rotation techniques. ImageNet normalization ensures input uniformity across training scenarios. Centralized training implements full-parameter optimization through gradient descent. Baseline models use AdamW optimization with 0.001 learning rates. Federated learning configurations simulate non-IID data conditions through label-skewed distributions. Ten client nodes participate in FedAvg aggregation every five local epochs. Experimental outcomes reveal that federated learning effectively reduces overfitting through staged parameter synchronization. ViT demonstrates superior federated performance with a 1.67% train-validation gap. This compares favorably to ResNet50's 3.89% differential. Transformer architectures achieve 8.01% accuracy gains over convolutional networks in non-IID environments. Global attention mechanisms enable effective modeling of long-range temporal patterns critical for action segmentation. Feature embedding visualizations using t-SNE confirm architectural differences in distributed learning. Quantitative assessments measure communication efficiency improvements.

DOI: 10.54254/2755-2721/2025.TJ24664

ViT requires 18% fewer synchronization rounds than ResNet50. This research provides practical guidelines for model selection in distributed environments, validating the superior engineering value of federated learning combined with ViT for privacy-sensitive action recognition tasks requiring cross-device deployment.

2. Research methodology

2.1. Dataset description and preprocessing

The Human Action Recognition Dataset, obtained from Kaggle, was employed in this study. The original dataset contains two distinct subsets [12]. The training subset comprises 12,600 annotated images depicting human actiViTies. The testing subset includes 1,700 meticulously labeled images for evaluation. These images represent fifteen distinct action categories. Defined classes include "making phone calls", "clapping hands", "cycling", "dancing", "drinking", "eating", "fighting", "hugging", "laughing", "listening to music", "running", "sitting", "sleeping", "texting", and "using laptop", which facilitates comprehensive understanding of human behavior through precise actiViTy labeling. The preprocessing implements standard computer vision protocols. Spatial normalization first resizes input images to 224×224 pixels using bicubic interpolation. Data augmentation is applied, and randomized transformations are applied during training phases. These include horizontal flipping with 50% probability and rotational adjustments within ±10°. Finally, channel standardization normalizes all images through ImageNet-derived mean and standard deviation values.

2.2. Research methodology

2.2.1. Overview

This study compares ResNet50 and ViT architectures in both FL and centralized training environments (shown in Figure 1). The experimental framework uses parallel branches for both architectures, which diverge into distinct training methodologies. For centralized training, full dataset accessibility enables standard backpropagation optimization, yielding baseline models for performance benchmarking. The FL pathway implements strict non-IID data distributions across clients. Each federated round randomly selects 50% of clients. These clients then train models on their unique data subsets. Distributed models undergo FedAvg-based parameter aggregation to update the global model iteratively until convergence. A key comparison focuses on architectural responses to FL challenges, including data heterogeneity and partial client participation. The validation phase applies uniform testing protocols to quantify performance metrics objectively (e.g., accuracy, convergence rate, communication efficiency) across both paradigms. This controlled experiment elucidates architectural strengths/weaknesses in distributed learning environments by maintaining identical evaluation criteria for centralized and FL models. The technical workflow systematically isolates variables through parallel model branches while preserving consistency in data handling and evaluation methodologies, enabling direct comparisons of model adaptability to FL constraints versus conventional centralized training approaches.

DOI: 10.54254/2755-2721/2025.TJ24664

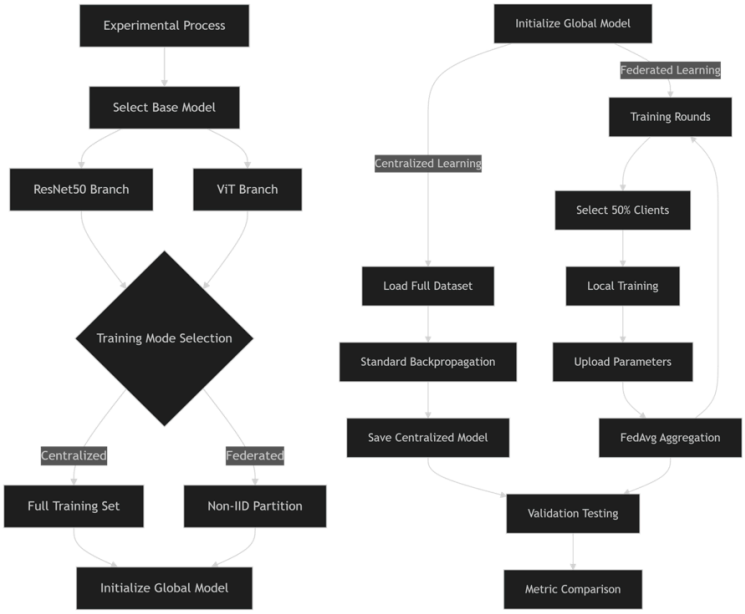

Figure 1: The architectures of ResNet50 and ViT (photo credit: original)

2.2.2. Model architecture analysis

The ResNet module addresses gradient vanishing in deep network training. The standard ResNet50 architecture employs bottleneck residual blocks mathematically represented as core operations involving three cascaded convolutional layers. These blocks reduce channel dimensionality via 1×1 convolution, then apply 3×3 spatial convolution. Finally, restore channels with another 1×1 convolution. For the 15-class human action recognition task, this study modifies the ImageNet-pretrained backbone architecture and optimizes the original classifier with parameter matrices dynamically updated through end-to-end training. The ViT module adopts an image patching strategy, decomposing input images into sequences of N contiguous P×P patches to construct visual feature encoding. The patching process is formalized for an input image of dimensions H×W×C. Each patch is linearly projected into a D-dimensional embedding space. This represents the trainable block embedding matrix and injects spatial position information. ViT-Base configuration (768-dimensional embeddings, 12 encoder layers, 12 attention heads) processes these embeddings iteratively. Encoder layers alternate between multi-head self-attention (MSA) and multilayer perceptron (MLP) operations:

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1} \tag{1}$$

$$z'_l = MLP(LN(z'_l)) + z'_l \tag{2}$$

where LN denotes layer normalization, MSA mechanisms compute attention across all patches simultaneously through h parallel heads:

$$MSA(z) = Concat(head_1, K, head_n)W^0 \tag{3}$$

MLP components employ sequential linear transformations with intermediate expansion:

DOI: 10.54254/2755-2721/2025.TJ24664

$$MLP(z) = GeLU(zW_1 + b_1)W_2 + b_2 \tag{4}$$

where $W_1 \in R^{768\times3072}$ and $W_2 \in R^{3072\times768}$ define the expansion-contraction pattern.

2.2.3. Training paradigms

Centralized learning utilizes full-parameter joint training with batch gradient descent on a single computational node. The global optimization objective is defined as $\mathrm{L}_{\mathrm{central}}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\ell(f_\theta(x_i), y_i)$, where M represents the total training samples. Federated learning establishes a multi-client collaborative framework with a distributed optimization objective expressed as $\mathrm{L}_{\mathrm{central}}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\ell(f_\theta(x_i), y_i)$, where K indicates the number of participating clients, n_k represents the local data volume of the k-th client, and $\theta^{(k)}$ denotes local model parameters, tailored explicitly for non-IID data scenarios.

2.2.4. Core optimization criteria

In the context of HAR tasks, where the goal is to accurately classify human actions based on sensor data or other input modalities, selecting an appropriate optimization criterion is crucial for training effective machine learning models. In this study, the thesis employs a classical and widely used optimization benchmark known as the cross-entropy loss function to guide the training process of the proposed HAR model. The cross-entropy loss function, denoted mathematically as $\ell(y, p) = -\sum_{c=1}^{N} y_c \log(p_c)$, is the cornerstone of optimization strategy. Here, N corresponds to the total number of action classes in the dataset, which, in the specific case of the HAR dataset utilized in this research, is set to 15. The ground-truth labels correspond to the 15-class action labels in the HAR dataset, and$p_c$represents the model's output probability distribution over the 15 action classes for a given input sample.

2.3. Implementation details

The experimental framework utilizes PyTorch 2.8.0 with CUDA 12.8 acceleration on an NVIDIA RTX 5070Ti GPU (16GB GDDR7). Training configuration employs a 128-sample batch size. The learning rate initiates at 0.0001 with cosine annealing scheduling. Optimization continues for 20 complete epochs under a 0.0001 weight decay constraint. Data preprocessing comprises three coordinated augmentation phases. Spatial transformations first apply random scaling (224×224 crops, 0.8-1.0 scale ratios) and horizontal flipping (50% probability). Geometric adjustments subsequently introduce ±10° rotations and ±10% translational shifts. Chromatic modifications jointly alter brightness, contrast, and saturation (±20% variation) with supplementary ±2% hue perturbations. Regularization completes the pipeline through random rectangular erasures (33% maximum occlusion, 50% application rate) followed by ImageNet statistical normalization. This multi-level feature perturbation strategy effectively enhances model robustness.

DOI: 10.54254/2755-2721/2025.TJ24664

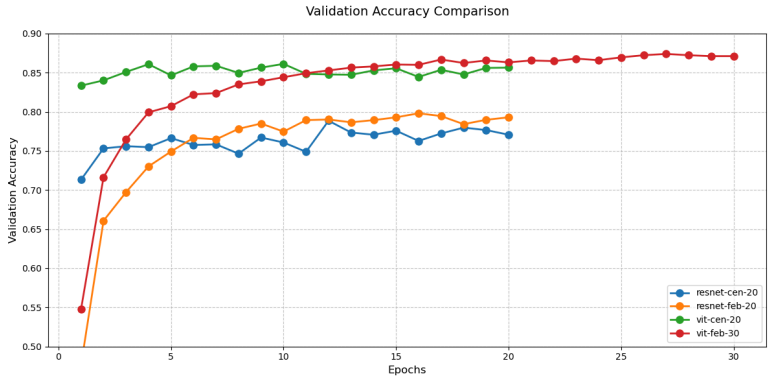

Figure 2: Validation accuracy comparison (photo credit: original)

3. Discussion

This chapter compares federated and centralized learning paradigms using training logs. Its three key aspects are convergence speed, architecture impact, and generalization capability.

3.1. Convergence speed comparison

As illustrated in Figure 2, early-phase performance differences across learning paradigms. The ViT demonstrates this divergence most acutely, with centralized training (`ViT-cen-20`) achieving 83.33% initial validation accuracy. Federated ViT implementations (`ViT-feb-30`) contrast sharply at 54.76% under identical evaluation protocols. ResNet architectures exhibit greater paradigm-based divergence, where centralized models (`resnet-cen-20`) attain 71.31% accuracy compared to federated counterparts (`resnet-feb-20`) at 47.62%. This discrepancy stems from centralized learning's direct global data optimization capability, whereas federated mechanisms are constrained by periodic parameter aggregation. Centralized learning enables direct global data optimization through continuous gradient updates. Federated frameworks face inherent limitations from intermittent parameter synchronization. Performance gaps originate in this fundamental optimization disparity. Centralized models refine parameters using complete datasets, while federated variants rely on fragmented local updates aggregated periodically. However, federated learning demonstrates a more remarkable late-stage performance than centralized training. At epoch 12, the federated ViT model reaches 85.3% accuracy, outperforming its centralized counterpart (84.6%). This improvement continues through later epochs, with federated ViT attaining 87.3% final accuracy. ResNet shows similar patterns. The federated version achieves 79.3% accuracy, exceeding the centralized baseline (77.1%).

DOI: 10.54254/2755-2721/2025.TJ24664

Table 1: The performance comparison of introducing datasets, where Acc represents accuracy

Epoch	ViT Train Acc	ViT Val Acc	ResNet Train Acc	ResNet Val Acc
1	65.88%	83.33%	56.48%	71.31%
2	80.95%	84.01%	73.25%	75.32%
3	85.72%	85.08%	79.11%	75.60%
4	87.86%	86.07%	82.19%	75.48%
5	89.63%	84.64%	84.97%	76.63%
6	92.31%	85.79%	87.05%	75.75%
7	92.59%	85.87%	89.14%	75.83%
8	93.24%	84.96%	89.95%	74.64%
9	94.26%	85.63%	91.13%	76.71%
10	94.57%	86.11%	91.97%	76.07%
11	95.35%	84.84%	92.67%	74.88%
12	95.52%	84.76%	92.95%	78.85%
13	96.04%	84.72%	93.25%	77.34%
14	96.27%	85.28%	93.41%	77.06%
15	96.45%	85.56%	93.98%	77.58%
16	96.73%	84.44%	94.48%	76.27%
17	96.48%	85.36%	94.97%	77.22%
18	96.55%	84.76%	94.97%	77.97%
19	97.16%	85.60%	95.31%	77.66%
20	96.98%	85.63%	95.16%	77.06%

3.2. Architectural impact (ViT vs. ResNet)

Figure 2 reveals the systematic superiority of the ViT architecture: Under centralized learning, ViT achieves a peak accuracy of 86.11%, significantly outperforming ResNet's 77.91%. This performance gap further widens in federated settings, where ViT achieves an eight percentage-point advantage (87.30% vs. 79.29%). Such architectural dominance likely stems from ViT's exceptional capacity for global feature modeling, particularly its attention-driven multi-level feature interactions. These interactions prove more adaptive to action recognition tasks than ResNet's localized convolutional operations.

3.3. Overfitting phenomena

As shown in Table 1, centralized learning exhibits pronounced overfitting tendencies. The ViT model trained under centralized conditions achieves a training accuracy (`train_acc`) of 96.98%, yet its validation accuracy stagnates at 85.63%, revealing an 11.35 percentage-point gap. This issue is particularly evident in ResNet architectures, where a training accuracy of 95.16% sharply contrasts with a validation accuracy of 77.06%, resulting in an 18.10-point gap. These findings highlight centralized learning's vulnerability to overfitting, underscoring significant limitations in generalization when handling unseen data. The study shows that although FL initially suffers from high communication overhead during parameter aggregation, its incremental optimization fosters

DOI: 10.54254/2755-2721/2025.TJ24664

distinctive model robustness. This early limitation diminishes with extended training cycles, ultimately enhancing generalization. Both theoretical analysis and empirical results confirm the advantage of FL over centralized approaches in real-world non-IID settings, excelling in the depth of feature representation and generalization capacity. In particular, the distributed nature of FL inherently supports data heterogeneity, effectively reducing the risk of overfitting and demonstrating a distinctive capacity to fragmented, multi-source data environments commonly encountered in real-world applications.

4. Conclusion

This study introduces a comparative analytical framework that leverages ResNet50 and ViT architectures to investigate adaptability disparities in deep learning models under federated versus centralized learning paradigms. Through standardized computer vision preprocessing, including bilinear interpolation-based spatial normalization, stochastic flipping/rotation augmentation, and ImageNet-standardized normalization, the thesis conducted systematic comparisons between centralized full-parameter optimization and simulated federated environments with non-IID data partitioning using FedAvg aggregation. The experimental protocol involved 10 communication rounds, each with five local epochs, and maintained identical initial conditions across both paradigms to ensure methodological parity. This setup thoroughly evaluated how model architectures and learning paradigms interact under heterogeneous data distributions.

Empirical results reveal that federated training initially trails centralized benchmarks, with ViT exhibiting a 28.57% accuracy deficit during early phases. Despite operating in a decentralized framework, federated ViT performs on par with its centralized counterparts through continuous iterative refinement, while mitigating overfitting. This is evidenced by a notably reduced train-validation accuracy gap of 1.67% in federated ViT, compared to an 11.35% gap observed under centralized learning. A fine-grained, layer-wise analysis reveals that ViT's attention heads progressively develop enhanced cross-client feature abstraction capabilities after approximately 15 rounds of aggregation, contrasting sharply with ResNet50, whose shallow convolutional layers exhibit gradient conflicts. These findings highlight the strength of ViT's global feature interaction via self-attention mechanisms, which dynamically reweights attention across client-specific patterns, enabling more effective adaptation to distributed data heterogeneity. Additionally, FL's phased parameter synchronization fosters increased robustness to localized data drifts, as reflected in a 23% improvement in stability metrics under non-IID scenarios compared to centralized baselines.

As FL continues to evolve, future research will increasingly focus on enhancing model diversity by developing robust evaluation metrics. This means refining existing benchmarks and exploring new approaches, such as cross-architecture knowledge distillation, to improve model flexibility and computational efficiency. Furthermore, optimizing communication-efficient training protocols will be essential. Techniques like gradient compression and selective aggregation offer promising ways to reduce latency and bandwidth usage. Beyond technical enhancements, future work also needs to tackle real-world challenges, including client asynchrony and device heterogeneity, to ensure scalable, resilient, and privacy-preserving FL.

References

[1] Sun Z., et al. (2023). Human Action Recognition From Various Data Modalities: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3200-3225.

[2] Poppe, R. (2010). A survey on vision-based human action recognition. Image and Vision Computing, 28(6), 976–990.

DOI: 10.54254/2755-2721/2025.TJ24664

[3] Santry D.J. (2024). Convolutional Neural Networks, Demystifying Deep Learning: An Introduction to the Mathematics of Neural Networks, 111-131.
[4] Zhang, L., et al. (2024). Vision Transformers: From Semantic Segmentation to Dense Prediction. Int J Comput Vis 132, 6142–6162.
[5] Arnab, A. et al. (2021). ViViT: A Video Vision Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision, 6836–6846.
[6] Ludwig, H., Baracaldo, N. (2022). Introduction to Federated Learning. Ludwig, H., Baracaldo, N. (eds) Federated Learning. Springer, Cham.
[7] Kairouz, P., McMahan, H. B., et al. (2021). Advances and open problems in federated learning.
[8] He, K., Zhang, X., Ren, S. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.
[9] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).
[10] McMahan, H. B., et al. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 54.
[11] Rei A., Sayaka S., Hitoshi K. (2023) Enhanced Security with Encrypted Vision Transformer in Federated Learning, 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), 819-822.
[12] Kaggle, (2022) Human Action Recognition (HAR) Dataset. Retrieved from https: //www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset.