연합학습의 Non-IID 환경에서 CNN-VAE와 ViT-VAE 모델의 잠재 표현 전이 효과 비교

(Comparison of Latent Representation Transfer Effects between CNN-VAE and ViT-VAE in Non-IID Federated Learning Environments)

박지우*, 강태욱*, 박철영**, 신창선***

(Ji Woo Park, Tae Wook Kang, Chul Young Park, Chang Sun Shin)

요 약

본 논문은 연합학습의 극심한 데이터 불균형(Quantity Skew Non-IID) 환경에서 발생하는 저자원 클라이언트의 성능 저하를 해결하기 위해 **'2단계 VAE 잠재 표현 전이 프레임워크'**를 제안한다. Phase 1에서 글로벌 VAE를 사전 학습하여 데이터 분포 지식을 잠재 공간에 인코딩하고, Phase 2에서 가중치를 통해 잠재 벡터를 분류기에 융합한다. CIFAR-100 및 Tiny-ImageNet 실험 결과, Baseline 대비 유의미한 성능 향상을 달성했으며, CNN과 ViT 아키텍처 간 통계적 동등성을 확인했다. 특히, 연합학습 특성상 인코더 동결이 글로벌 표현 왜곡을 방지하는 필수 설계 원칙임을 규명했다. 본 프레임워크는 기존 알고리즘(FedAvg, FedProx 등)과 직교하며, 라운드당 통신 효율성도 그대로 유지한다.

■ 중심어 : 연합학습 ; Non-IID ; Variational Autoencoder ; Vision Transformer ; 잠재 표현 전이

Abstract

This paper proposes a two-stage VAE latent representation transfer framework to mitigate the performance degradation of resource-constrained clients in Federated Learning under severe data imbalance (Quantity Skew Non-IID) environments. In Phase 1, a global VAE is pre-trained to encode knowledge of the entire data distribution into a latent space. In Phase 2, the latent vectors are fused into the classifier via learnable weights. Experimental results on CIFAR-100 and Tiny-ImageNet demonstrate significant performance gains over the baseline and confirm statistical equivalence between CNN and ViT architectures. Notably, we experimentally reveal that freezing the encoder is an essential design principle to prevent global representation distortion caused by Non-IID data. This framework is orthogonal to conventional model aggregation algorithms (e.g., FedAvg, FedProx) and maintains communication efficiency by keeping the per-round communication cost identical to single-stage federated learning.

■ keywords : Federated Learning ; Non-IID ; Variational Autoencoder ; Vision Transformer ; Latent Representation Transfer

Ⅰ. 서 론

최근 딥러닝은 대규모 데이터셋과 고성능 연산 자원을 바탕으로 비약적인 발전을 이루었으나, 의료·금융 등 데이터 프라이버시가 중요한 분야에서는 중앙 집중식 학습의 한계에 부딪히고 있다[1,2]. 이를 해결하기 위해 제안된 연합학습(Federated Learning)은 로컬 데이터를 전송하지 않고 모델 파라미터만을 공유함으로써 프라이버시를 보존한다[3]. 그러나 현실적인 연합학습 환경에서는 클라이언트 간 데이터 분포가 상이한 Non-IID(Non-Independent and Identically Distributed) 문제가 발생하며, 특히 클라이언트별 데이터 보유량의 차이가 극심한 Quantity Skew 환경은 저자원 클라이언트의 과적합과 일반화 성능 저하를 야기하는 핵심 난제로 지목된다[4].

기존 연구들이 주로 클래스 불균형(Label Skew)에

* 준회원, 국립순천대학교 정보통신공학전공 석사과정
** 정회원, 국립순천대학교 인공지능공학전공 교수
*** 종신회원, 국립순천대학교 인공지능공학전공 교수
이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00407739).

접수일자 : 2026년 04월 07일
수정일자 : 2026년 05월 14일
게재확정일 : 2026년 05월 25일
교신저자 : 박철영,신창선 e-mail :cypark@sonu.ac.kr, csshin@sonu.ac.kr

집중해 온 것과 달리, 본 연구는 Quantity Skew 환경에서 VAE(Variational Autoencoder)의 잠재 표현(Latent Representation) 전이를 활용한 성능 개선 방안을 제안한다. 제안하는 2단계 프레임워크는 먼저 연합학습을 통해 전체 데이터 분포를 반영하는 글로벌 VAE 인코더를 구축하고(Phase 1), 이를 분류 모델의 보조 특성으로 융합하여 저자원 클라이언트의 표현 학습을 보강한다(Phase 2).

특히 본 연구에서는 잠재 표현 추출의 핵심인 인코더 아키텍처에 주목하여, 지역적 특성에 강한 CNN-VAE와 글로벌 관계를 포착하는 ViT-VAE의 전이 효과를 체계적으로 비교 분석하였다.

Ⅱ. 관련 연구

1. 연합학습(Federated Learning)

연합학습의 가장 기본 알고리즘은 McMahan 등이 제안한 FedAvg(Federated Averaging)이다[3]. 서버가 글로벌 모델 파라미터를 선택된 클라이언트들에게 배포하면, 각 클라이언트는 로컬 데이터를 활용하여 정해진 에포크 동안 학습을 수행하고 업데이트된 모델을 생성한다. 이후 서버는 클라이언트들로부터 전달받은 모델을 취합하며, 각 데이터 크기에 비례하여 가중 평균하는 방식으로 글로벌 모델을 업데이트한다.

FedAvg는 단순하면서도 효과적이나, Non-IID 환경에서는 클라이언트 간 학습 방향의 불일치로 인해 수렴이 불안정해지는 클라이언트 드리프트(client drift) 현상이 발생한다[5]. 이를 완화하기 위해 Li 등이 제안한 FedProx는 로컬 학습 목적 함수에 근접 항을 추가한다[6].

$$\min_w F_{k(w)} + \frac{\mu}{2}|w - w_t|^2 \quad (1)$$

여기서 μ는 근접 항의 강도를 조절하는 하이퍼파라미터이며, wt는 현재 라운드의 글로벌 모델이다. 이 항은 로컬 모델이 글로벌 모델에서 과도하게 이탈하는 것을 막아 학습 안정성을 높인다. 이 외에도 FedMA는 뉴런 매칭 기반 집계를, MOON은 모델 대조 학습을 통해 Non-IID 문제를 완화하는 접근을 제시하였다[7,8]. 2022년 이후에도 글로벌 지식 보존을 위한 비참(non-true) 클래스 증류 기반 FedNTD[9], 이질적 환경에서의 표현 공간 차원 붕괴를 완화하는 FedDecorr[10], 데이터 이질성 대응을 위한 아키텍처 설계 관점의 분석[11] 등 활발한 후속 연구가 이어지고 있으나, 모두 모델 파라미터 공간 또는 학습 손실단의 개입에 머문다. 본 연구는 이러한 흐름과 달리 표현 공간(latent space)에서 글로벌 분포 지식을 사전 학습하여 분류기 입력 단에서 보강하는 직교적 접근을 취하므로 위 기법들과 결합 가능한 보완 모듈로 작동한다.

2. Variational Autoencoder (VAE)

VAE는 인코더-디코더 구조의 생성 모델로, 입력 이미지를 잠재 공간의 확률 분포로 매핑한다[12]. 인코더는 입력을 잠재 변수의 분포 파라미터(평균 μ, 분산 σ^2)로 변환하고, 디코더는 샘플링된 잠재 변수로부터 입력을 복원한다. VAE의 손실 함수는 재구성 손실(reconstruction loss)과 KL 발산(Kullback-Leibler divergence)의 합으로 구성된다.

$$L_{VAE} = L_{recon} + \beta \cdot D_{KL}(q_\phi(z|x)||p(z)) \quad (2)$$

재구성 손실은 입력과 복원 이미지 간의 차이를 측정하며, KL 발산은 학습된 잠재 분포가 표준 정규 분포 N(0, I)에 가까워지도록 정규화한다. 학습 시에는 잠재 변수의 샘플링 과정을 미분 가능하게 만들기 위해 재매개변수화를 적용한다. 또한, 학습 초기에 KL 발산 항이 과도하게 작용하여 의미 있는 표현 학습을 방해하는 KL 붕괴(KL collapse)를 방지하고자, KL 항의 가중치 β를 점진적으로 증가시키는 KL Annealing 기법을 활용한다[13].

3. Vision Transformer (ViT)

Dosovitskiy 등이 제안한 ViT(Vision Transformer)는 입력 이미지를 고정된 크기(P×P)의 패치(patch)로 분할한 후, 각 패치를 선형 투영하여 고정 차원의 임베딩 벡터 토큰으로 변환하는 패치 임베딩(Patch Embedding) 과정을 거친다. 변환된 토큰에는 공간적 위치 정보를 보존하기 위해 학습 가능한 위치 인코딩(Positional Encoding)이 더해져 Transformer 인코더

에 입력된다[14]. 인코더 내부의 핵심인 다중 헤드 셀프 어텐션(Multi-Head Self-Attention, MHSA)은 모든 패치 간의 상호작용을 동시에 모델링한다. 최종적으로 이 어텐션 모듈의 출력값은 다층 퍼셉트론(MLP Block)을 통한 비선형 변환을 거쳐 모델의 특징 표현을 고도화한다.

CNN과 ViT는 이미지 데이터를 처리하는 아키텍처적 특성에서 뚜렷한 대조를 이룬다. CNN은 지역성과 이동 불변성이라는 강한 귀납적 편향을 내재하여 소량의 데이터로도 효율적인 학습이 가능하지만 전역적 패턴 인식은 깊은 레이어에 도달해야만 가능하다. 반면, ViT는 귀납적 편향을 최소화한 데이터 의존적 구조를 취하여 첫 레이어부터 글로벌 패턴을 모델링할 수 있으나, 사전 학습이 전제되지 않으면 대규모 학습 데이터가 요구된다. 이러한 차이는 VAE 인코더로 사용될 때 잠재 표현의 특성에도 영향을 미치며, 잠재 표현 전이 효과의 차이로 이어진다.

4. 연합학습에서의 생성 모델 활용

Non-IID 환경에서 생성 모델을 활용한 연구는 크게 두 방향으로 분류된다. 첫째는 합성 데이터를 생성하여 부족한 데이터를 보충하는 데이터 증강 접근법으로, 생성 이미지의 품질 의존도와 프라이버시 우려가 존재한다[15,16]. 둘째는 본 연구에서 채택한 방식으로, VAE 인코더가 학습한 잠재 벡터를 분류기의 보조 특징으로 활용하는 표현 학습 기반 접근법이다. 이 방식은 이미지 재구성 품질에 대한 의존도를 낮추고, 글로벌 VAE를 통해 전체 데이터 분포의 지식을 잠재 공간에 인코딩하므로 개별 클라이언트의 편향된 데이터로 인한 성능 한계를 보완할 수 있다. 기존 연구는 주로 Label Skew 환경에 초점을 맞추고 있으나, 본 연구는 Quantity Skew 환경에 주목하여 두 가지 VAE 아키텍처(CNN vs ViT)의 전이 효과를 다중 데이터셋에서 비교 분석한다는 점에서 차별성을 갖는다[17,18].

III. 시스템 설계 및 구성

1. 전체 시스템 아키텍처

본 연구에서 제안하는 시스템은 2단계(Phase 1, Phase 2)로 구성된 연합학습 프레임워크로, Flower(FLWR) 시뮬레이션 환경에서 동작한다. 전체 구조는 그림 1과 같다.

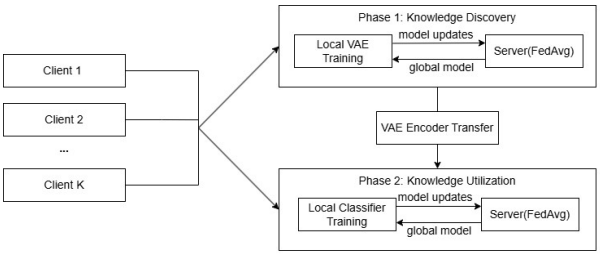

그림 1. 시스템 구성도

2. Non-IID 데이터 분할: Quantity Skew

본 연구에서 Non-IID 환경은 클라이언트 간 데이터 양의 불균형(quantity skew)으로 정의한다. 전체 데이터셋의 샘플을 K개 클라이언트에 배분하기 위해, Dirichlet 분포 Dir(α)에서 비율 벡터 p를 샘플링한다.

$$p \sim Dir(\alpha,\ldots,\alpha) \tag{3}$$

$$n_k = \max(\lfloor p_k \cdot N \rfloor, n_{\min}) \tag{4}$$

여기서 α는 불균형 정도를 제어하는 파라미터로, 값이 작을수록 불균형이 심해지며 $\alpha \to \infty$이면 균등 분배에 수렴한다. 본 연구에서는 α = 0.3으로 설정하여 현실적인 수준의 quantity skew를 모사하고, 최소 샘플 수는 32로 적용하였다. 실험 분석을 위해 클라이언트를 보유 데이터 양 기준으로 Low-resource(하위 25%), Mid-resource(중간 50%), High-resource(상위 25%) 세 그룹으로 분류하여 잠재 표현 전이의 효과를 세분화하여 분석한다.

그림 2, 3는 α = 0.3 조건에서 시드 42로 생성된 10개 클라이언트의 데이터 배분 결과를 데이터셋별로 나타낸 것이다. Quantity Skew는 클래스 레이블의 분포 자체는 동일하게 유지하되 클라이언트별 보유 샘플 수에 극심한 차이를 유발하는 방식으로, 일부 클라이언트는 전체 데이터의 20% 이상을 보유하는 반면 소수 클라이언트는 최소 샘플 수(32개) 수준에 그친다. 이러한 설정은 데이터 접근성이 불균등한 실제 연합학습 시나리오를 반영한다.

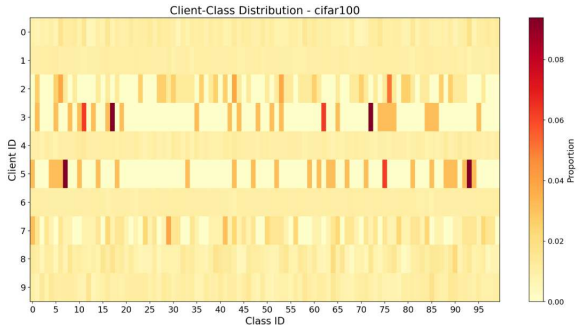

그림 2. 클라이언트별 데이터 배분 분포—CIFAR-100

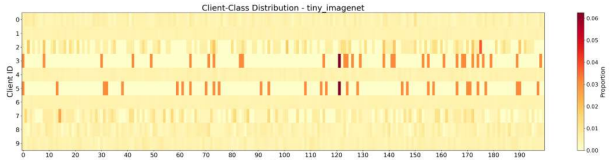

그림 3. 클라이언트별 데이터 배분 분포—Tiny-ImageNet

3. Phase 1: VAE 연합 사전 학습

Phase 1의 목적은 개별 클라이언트의 편향된 로컬 데이터만으로는 얻기 어려운, 전체 데이터 분포를 반영하는 글로벌 VAE를 구축하는 것이다. 본 연구에서는 두 가지 인코더 아키텍처를 비교한다. CNN-VAE의 인코더는 4개의 합성곱 블록(Conv2d-BatchNorm-ReLU, stride=2)으로 공간 해상도를 점진적으로 축소하고, 최종 특성 맵을 평탄화한 후 완전 연결 레이어를 통해 잠재 분포의 평균 μ와 로그 분산 $\log \sigma^2$(각 128차원)를 출력한다. ViT-VAE의 인코더는 입력 이미지를 P×P 패치로 분할하여 선형 투영하고, 학습 가능한 위치 임베딩을 추가한 후 여러 개의 Transformer 블록과 LayerNorm을 통과시킨다. 이후 모든 패치 임베딩에 대한 글로벌 평균 풀링(Global Average Pooling)으로 고정 크기 벡터를 생성하여 평균 μ와 로그 분산 $\log \sigma^2$를 출력한다. 두 모델의 디코더는 각각 전치 합성곱(ConvTranspose2d)과 MLP+전치 합성곱을 사용하여 잠재 벡터로부터 이미지를 복원한다. 주요 하이퍼파라미터로 잠재 차원 128, ViT의 경우 embed_dim=192, depth=4, num_heads=3, patch_size=8을 사용하며, KL Annealing은 전체 라운드의 50%까지 선형 워밍업한다. VAE 연합 학습은 FedAvg를 사용하여 모든 클라이언트가 매 라운드 참여한다.

표 1. VAE 기본 하이퍼파라미터

파라미터	값	설명
잠재차원	128	잠재 벡터의 차원 수
KL 가중치	1.0	KL Annealing 최종 값
KL Annealing	선형 워밍업	전체 라운드의 50%까지 0-> 1
ViT embed_dim	192	Transformer 임베딩 차원
ViT depth	4	Transformer 블록 수
ViT num_heads	3	Multi-head attention 헤드 수
ViT patch_size	8	패치 크기

4. Phase 2: 잠재 표현 전이 기반 분류기 학습

Phase 2에서는 네 가지 모드의 분류기를 학습하고 비교한다. Baseline은 4개의 합성곱 블록(Conv2d - BatchNorm - ReLU-MaxPool)과 Adaptive Average Pooling(2×2)으로 1024차원(256채널 × 2 × 2) 특성 벡터를 추출하는 경량 CNN 분류기이다. CNN-VAE/ViT-VAE 모드는 이 Baseline 백본에 사전 학습된 VAE 인코더의 잠재 벡터를 학습 가능한 가중치로 결합하는 이중 경로 융합 구조이다. 입력 이미지는 두 경로를 동시에 통과한다. 첫 번째 경로에서는 CNN 백본이 이미지를 처리하여 1024차원의 백본 특성 벡터를 추출하고, 두 번째 경로에서는 데이터셋 정규화를 역변환한 이미지가 사전 학습된 VAE 인코더에 입력되어 128차원의 잠재 벡터를 생성한다. 이후 잠재 벡터에 학습 가능한 가중치 α = sigmoid(α_logit)를 곱하여 기여도를 조절한 뒤, 백본 특성 벡터와 연결(concatenate)하면 1152차원의 융합 표현이 된다. 이 융합 표현은 Linear(1152→512)—LayerNorm—ReLU—Dropout(0.3)—Linear(512→클래스 수)로 구성된 분류 헤드를 거쳐 최종 로짓을 출력한다.

α_logit은 학습 가능한 스칼라 파라미터로, 초기값 -1.4(α ≈ 0.20)에서 시작하여 백본 특성을 우선시하면서 잠재 벡터의 기여도를 점진적으로 조절한다. Phase 2에서 분류기 입력은 데이터셋별 평균/표준편차로 정규화되어 있으나, VAE는 [0, 1] 범위 이미지로 학습되었으므로 VAE 인코더 입력 전에 다음과 같이 역정규

화한다. VAE 인코더의 파라미터 처리는 동결(Freeze)과 비동결(Unfreeze) 두 전략을 비교하며, 동결 전략은 글로벌 분포 지식을 보존하고 비동결 전략은 태스크 특화 표현 학습을 도모하나 Non-IID 환경에서 글로벌 표현 훼손의 위험이 존재한다.

5. 학습 안정화 기법

연합학습의 Non-IID 환경에서 학습 안정성을 확보하기 위해 세 가지 기법을 적용한다. 첫째, FedProx의 근접 항($\mu = 0.01$)을 로컬 학습 목적 함수에 추가하여 글로벌 모델로부터의 과도한 이탈을 억제한다. 둘째, 라운드 진행에 따라 로컬 학습률을 코사인 스케줄로 감소시키되 최소 학습률을 초기값의 10%로 설정한다. 셋째, 그래디언트 노름 클리핑($\tau = 10.0$)으로 그래디언트 폭발을 방지한다.

6. 실험 설정

본 연구에서는 CIFAR-100(32×32, 100 클래스, 50K 샘플)과 Tiny-ImageNet(64×64, 200 클래스, 100K 샘플) 두 데이터셋을 사용한다. 실험 환경은 아래 표 2와 같다.

표 2. 실험 환경

항목	사양
GPU	NVIDIA GeForce RTX 5080(16GB VRAM)
CPU	Intel Core i9-14900
프레임워크	PyTorch, Flower(FLWR)
시뮬레이션	Ray 기반 가상 클라이언트
클라이언트 수	10
라운드 수	Phase 1: 20, Phase 2: 30
로컬 에포크	3
배치 크기	64
학습률	0.001
시드	42

주요 하이퍼파라미터와 그 설정 근거는 다음과 같다. Phase 1의 20 라운드는 VAE 재구성 손실이 충분히 감소하는 시점으로, 분류 목적의 Phase 2보다 적은 라운드로 안정적인 잠재 공간 구축이 가능하기 때문이다. Phase 2의 30 라운드는 수렴을 보장하는 동시에 실험의 현실적 소요 시간을 관리하는 균형점으로 결정하였으며, CIFAR-100에서는 Round 20 이후 수렴이 확인되어 충분한 여유를 확보한다. 로컬 에포크 3은 FedAvg 문헌의 권고 범위(1~5) 내에서, Non-IID 환경의 클라이언트 드리프트를 억제하기 위해 보수적으로 설정하였다. 배치 크기 64와 학습률 0.001(Adam 옵티마이저)은 딥러닝 표준 범위 내에서 GPU 메모리 효율과 학습 안정성을 균형 있게 고려한 값이며, 시드 42는 데이터 분할 및 모델 초기화의 재현성을 보장한다. 평가 체계는 세 수준으로 구성된다. 글로벌 수준에서는 전체 테스트셋에 대한 Accuracy와 Macro F1-Score를, 클라이언트 수준에서는 데이터 보유량 기준 Low/Mid/High 리소스 그룹별 평균 및 표준편차를, 학습 과정 수준에서는 라운드별 손실 및 정확도를 추적한다. 정밀도(Precision)와 재현율(Recall)을 별도 보고하지 않은 것은, Macro F1-Score가 이 두 지표의 조화 평균을 클래스 간 균등 가중치로 집계하므로 100~200개 클래스 환경에서 단일 요약 지표로 충분하기 때문에 제외하였다.

Ⅳ. 실험 및 결과

1. 실험 개요

본 장에서는 제안하는 2단계 연합학습 프레임워크의 성능을 다양한 조건에서 체계적으로 평가한다. 실험은 크게 세 가지 축으로 구성된다. 실험 A (핵심 비교), Baseline, CNN-VAE, ViT-VAE의 인코더 동결 조건에서의 성능 비교. CIFAR-100(32×32, 100 클래스)과 Tiny-ImageNet(64×64, 200 클래스) 두 데이터셋에서 수행하였다. 실험 B (인코더 비동결), VAE 인코더의 파라미터를 Phase 2에서 함께 미세 조정하는 비동결(unfreeze) 조건에서의 성능 변화를 분석한다. 모든 실험에서 Non-IID 환경은 Dirichlet 분포($\alpha = 0.3$)를 사용한 Quantity Skew로 설정하였으며, 10개 클라이언트, Phase 1은 20 라운드, Phase 2는 30 라운드로 진행하였다. 시드 42로 재현성을 확보하였다.

2. 실험 A: 핵심 성능 비교

표 3은 두 데이터셋에서의 최종 성능(Round 30)을 요약한다.

표 3. Phase 2 글로벌 성능 비교 (Round 30, 인코더 동결)

모드	CIFAR-100 Acc	CIFAR-100 F1	Tiny-ImageNet Acc	Tiny-ImageNet F1
Baseline	60.18%	0.5989	38.85%	0.3728
CNN-VAE	61.73%	0.6150	44.29%	0.4319
ViT-VAE	61.68%	0.6140	44.77%	0.4361
Δ(CNN-VAE)	+1.55%p	+0.0161	+5.44%p	+0.0591
Δ(ViT-VAE)	+1.50%p	+0.0151	+5.92%p	+0.0633

두 VAE 모델 모두 Baseline 대비 유의미한 성능 향상을 달성하였다. 특히 Tiny-ImageNet에서의 향상폭(+5.4~5.9%p)이 CIFAR-100(+1.5~1.6%p)보다 현저히 크다. 이는 200개 클래스의 복잡한 분류 태스크에서 잠재 표현이 제공하는 보조 정보의 가치가 더 높기 때문으로 해석된다. 또한 CNN-VAE는 CIFAR-100(32×32)에서 ViT-VAE는 Tiny-ImageNet (64×64)에서 각각 상위 성능을 기록하여 해상도 의존적 아키텍처 우위가 관찰되었다. 이는 CNN-VAE의 합성곱 연산이 저해상도 이미지의 지역적 특성을 효율적으로 추출하는 반면, ViT-VAE의 self-attention은 고해상도 이미지에서 풍부한 패치 간 글로벌 관계를 포착하여 우위를 보이기 때문이다. 32×32 이미지에서는 패치 수가 16개로 제한되어 self-attention의 이점이 감소한다.

표 4는 보유 데이터 양에 따라 Low(하위 25%), Mid(중간 50%), High(상위 25%)로 분류한 클라이언트 그룹별 성능이다.

표 4. CIFAR-100 클라이언트 그룹별 Accuracy

그룹	Baseline	CNN-VAE	ViT-VAE
Low	63.5%±0.04	68.3%±0.27	84.1%±0.14
Mid	61.2%±0.05	64.4%±0.04	64.4%±0.02
High	58.8%±0.01	60.4%±0.01	60.9%±0.00

표 5. Tiny-ImageNet 클라이언트 그룹별 Accuracy

그룹	Baseline	CNN-VAE	ViT-VAE
Low	17.8%±0.14	53.3%±0.33	37.8% ± 0.22
Mid	38.7%± 0.01	41.1%±0.04	42.1%±0.01
High	38.0%±0.01	43.3%±0.00	44.1%±0.01

주목할 점은 저자원 클라이언트(Low 그룹)에서의 성능 향상이 가장 두드러진다는 것이다. CIFAR-100 Low 그룹은 Baseline 63.5%에서 ViT-VAE 84.1%로 +20.6%p 향상되었고, Tiny-ImageNet Low 그룹은 Baseline 17.8%에서 CNN-VAE 53.3%로 +35.5%p 향상되었다. 이 결과는 "글로벌 VAE의 잠재 표현이 저자원 클라이언트의 표현 학습을 보강한다"는 본 프레임워크의 핵심 가설을 강력하게 지지한다.

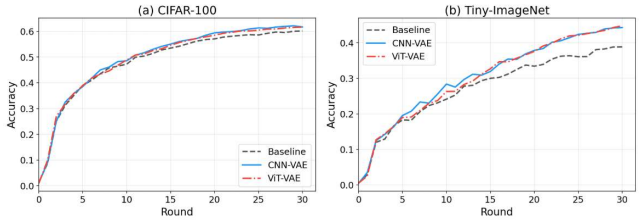

그림 4. Round 별 Accuracy

그림 4는 라운드별 글로벌 정확도 변화를 보여준다. 두 데이터셋 모두에서 VAE 모델은 Baseline보다 빠르게 수렴하며, 특히 학습 초기(Round 1~10) 구간에서의 차이가 크다. 이는 사전 학습된 잠재 표현이 분류기의 초기 학습 방향을 안내하는 역할을 함을 시사한다. CIFAR-100에서는 Round 20 이후 세 모델 모두 수렴 단계에 진입하며, 정확도 변동이 ±0.5%p 이내로 안정적이다. Tiny-ImageNet에서는 Round 30까지도 미세한 상승 추세가 관찰되어, 더 복잡한 태스크에서는 추가 라운드가 유효할 수 있음을 시사한다. Phase 2 학습 완료 후 수렴된 α = sigmoid(α_logit) 값은 다음과 같다.

표 6. 학습된 α 값

데이터셋	CNN-VAE α	ViT-VAE α
CIFAR-100	0.198	0.203
Tiny-ImageNet	0.228	0.227

모든 조건에서 α ≈ 0.20~0.23으로 수렴하여, 잠재 벡터가 백본 특성의 약 20%에 해당하는 가중치로 융합됨을 확인하였다. 이는 분류의 주요 정보는 CNN 백본이 추출하고 잠재 벡터는 보완적 역할을 수행함을 의미한다. Tiny-ImageNet (0.228)이 CIFAR-100 (0.198)보다 약간 높아 복잡한 태스크에서 잠재 표현의 기여도가 증가하며, CNN-VAE와 ViT-VAE의 alpha 값이 거의 동일한 것은 융합 가중치가 인코더 아키텍처가 아닌 태스크 특성에 의해 결정됨을 나타낸다.

3. 실험 B: 인코더 동결 vs 비동결

VAE 인코더의 파라미터를 Phase 2에서 함께 미세 조정(fine-tuning)하는 비동결 조건의 효과를 분석하였다.

표 7. 인코더 동결(Freeze) vs 비동결(Unfreeze) 성능 비교

모드	CIFAR-100 Acc	CIFAR-100 F1	Tiny-ImageNet Acc	Tiny-ImageNet F1
CNN-VAE Freeze	61.73%	0.6150	44.29%	0.4319
CNN-VAE Unfreeze	61.88%	0.6176	44.29%	0.4312
ViT-VAE Freeze	61.68%	0.6140	44.77%	0.4361
ViT-VAE Unfreeze	61.56%	0.6138	43.67%	0.4267

인코더 비동결은 전반적으로 성능 향상에 기여하지 못하며, ViT-VAE의 경우 오히려 유의미한 하락을 초래하였다. 특히 ViT-VAE는 CNN-VAE보다 비동결에 더 취약한데, Tiny-ImageNet에서 ViT-VAE의 비동결 성능 하락(-1.10%p)은 CNN-VAE(±0.00%p)와 대조적이다. 이는 ViT-VAE의 파라미터 수(3.09M)가 CNN-VAE(1.17M~ 2.36M)보다 크기 때문에 Non-IID 데이터로의 미세 조정 시 과적합 위험이 더 높기 때문이다. 중앙 집중식 학습에서는 인코더 미세 조정이 일반적으로 성능을 향상시키지만, 연합학습에서는 각 클라이언트의 편향된 로컬 데이터로 Phase 1의 글로벌 분포 표현이 왜곡되고, FedAvg로 서로 다른 방향으로 업데이트된 인코더를 단순 가중 평균하면 글로벌 표현의 품질이 저하되기 때문에 반대 결과가 관찰되었다.

Low 그룹에서 비동결의 영향이 가장 크게 나타난다. ViT-VAE Low 그룹은 84.1% → 68.3%로 15.8%p 하락하였다. 이는 저자원 클라이언트에서 인코더의 대량 파라미터가 소량 로컬 데이터에 과적합되어, Phase 1에서 획득한 글로벌 지식이 소실됨을 의미한다.

표 8. CIFAR-100 비동결 조건 그룹별 Accuracy

그룹	CNN-VAE Freeze	CNN-VAE Unfreeze	ViT-VAE Freeze	ViT-VAE Unfreeze
Low	68.3%	57.1%	84.1%	68.3%
Mid	64.4%	63.3%	64.4%	63.6%
High	60.4%	60.5%	60.9%	60.6%

4. 파라미터 효율성 분석

표 9 . VAE 인코더 파라미터 수 비교

VAE 타입	CIFAR-100	Tiny-ImageNet	배율
CNN-VAE	1.17M	2.36M	1.0×
ViT-VAE	3.09M	3.11M	1.3~2.6×

CNN-VAE는 ViT-VAE 대비 1.3~2.6배 적은 파라미터로 유사한 수준의 잠재 표현 전이 효과를 달성한다. 특히 CIFAR-100에서는 CNN-VAE가 2.6배 적은 파라미터로 오히려 더 높은 성능(61.73% vs 61.68%)을 기록하여, 저해상도 이미지에서의 파라미터 효율성이 돋보인다. 이는 실제 연합학습 배포 환경에서 중요한 고려 사항이다. 연합학습에서는 매 라운드마다 모델 파라미터를 클라이언트-서버 간 전송해야 하므로, 인코더 비동결 시 전송해야 할 파라미터 수가 통신 비용에 직접적으로 영향을 미친다. CNN-VAE의 작은 모델 크기는 통신 효율성 측면에서도 유리하다.

5. 종합 분석 및 논의

실험 결과를 종합하면, CNN-VAE와 ViT-VAE는 이미지 해상도에 따른 명확한 트레이드오프 관계를 보인다.

표 10. CNN-VAE vs ViT-VAE 종합 비교

평가 항목	CNN-VAE 우위	ViT-VAE 우위
저해상도 성능 (32×32)	✓	
고해상도 성능 (64×64)		✓
파라미터 효율성	✓	
비동결 안정성	✓	
Low 그룹 향상 (C100)		✓
Low 그룹 향상 (TI)	✓	

이러한 트레이드오프는 두 아키텍처의 근본적 차이에 기인한다. CNN의 지역적 귀납적 편향은 저해상도에서 제한된 정보를 효율적으로 압축하여 적은 파라미터로도 충분한 잠재 표현을 생성하는 반면, ViT의 글로벌 self-attention은 고해상도에서 풍부한 패치 간 관계를 포착하여 더 표현력 있는 잠재 벡터를 생성하지만 저해상도에서는 패치 수의 한계로 이점이 감소한다. 잠재 표현 전이가 효과적인 메커니즘은 다음과 같이 분석된다. Phase 1에서 모든 클라이언트의 데이

터로 학습된 VAE 인코더가 전체 데이터 분포를 반영하는 잠재 공간을 구축하여 Phase 2에서 글로벌 컨텍스트를 제공하고, CNN 백본(분류 목표)과 VAE 인코더(재구성 목표)의 이중 정보 경로가 상호 보완적 특성을 추출한다. 또한 학습 가능한 가중치(α ≈ 0.20)가 잠재 벡터의 기여를 적절히 제한하여 과적합 방지와 보조 정보 전달 간의 균형점을 자동으로 학습한다.

본 연구의 결과를 토대로, 연합학습 환경에 최적화된 잠재 표현 전이 가이드라인을 다음과 같이 제안한다. Non-IID 특성상 인코더는 반드시 동결해야 하며, 아키텍처 선택은 이미지 해상도에 따라 저해상도에서는 CNN-VAE, 고해상도에서는 ViT-VAE가 적합하다. 리소스 제약 환경에서는 2.6배 적은 파라미터로 유사한 성능을 달성하는 CNN-VAE를 우선하는 것이 바람직하다.

V. 결 론

본 연구에서는 연합학습의 Quantity Skew Non-IID 환경에서 CNN-VAE와 ViT-VAE의 잠재 표현 전이 효과를 비교 분석하기 위한 2단계 프레임워크를 제안하고 검증하였다. Phase 1에서 VAE를 연합학습으로 사전 학습하여 글로벌 잠재 공간을 구축하고, Phase 2에서 학습 가능한 가중치(learnable alpha)를 통해 잠재 표현을 분류기에 융합하는 방식이다.

CIFAR-100과 Tiny-ImageNet에서의 실험 결과, 두 VAE 모델 모두 Baseline 대비 유의미한 성능 향상(CIFAR-100: +1.5~1.6%p, Tiny-ImageNet: +5.4~5.9%p)을 달성하였으며, 특히 저자원 클라이언트에서 최대 +35.5%p의 향상을 보여 잠재 표현 전이가 데이터 불균형 환경의 성능 공정성 개선에 효과적임을 확인하였다. 아키텍처 측면에서는 CNN-VAE가 저해상도(32×32)에서, ViT-VAE가 고해상도(64×64)에서 각각 우위를 나타내어 귀납적 편향 차이에 따른 해상도 의존적 상호 보완성이 관찰되었다. 또한 연합학습 환경에서는 중앙 집중식 학습과 달리 인코더 비동결이 Non-IID 데이터에 의한 글로벌 표현 왜곡을 유발하여 성능을 저하시키며(ViT-VAE: -1.10%p), 인코더 동결이 필수적임을 규명하였다.

본 연구의 한계를 극복하고 발전시키기 위한 향후 연구 방향은 다음과 같다. Quantity Skew 외에 Label Skew, Feature Skew 등 다양한 Non-IID 유형과 Dirichlet α 파라미터 변화에 따른 성능 변화를 분석하는 심화 실험이 요구된다. 본 연구의 10개 클라이언트 환경을 50~100개 이상으로 확장하여 대규모 연합학습에서의 확장성과 통신 효율성을 검증하고, Hybrid CNN-Transformer, MLP-Mixer 등 최신 아키텍처를 VAE 인코더로 활용하는 방향도 고려할 수 있다. 또한 본 연구에서 ViT-VAE는 64×64 해상도의 Tiny-ImageNet에서만 CNN-VAE 대비 우위를 보였으나, ViT의 글로벌 self-attention이 본격적으로 효과를 발휘하는 128×128 이상의 고해상도 데이터셋으로 실험을 확장하여 해상도 증가에 따른 ViT-VAE의 잠재 표현 품질 변화를 체계적으로 검증할 필요가 있다. 마지막으로, 차분 프라이버시(Differential Privacy)나 안전한 집계(Secure Aggregation) 등의 기법을 적용하여 잠재 표현 전이가 추가적인 프라이버시 위험을 초래하지 않음을 검증하는 것이 필요하다.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436 - 444, May 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770 - 778, June 2016.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” *in Proc. of Artificial Intelligence and Statistics (AISTATS)*, pp. 1273 - 1282, Apr. 2017.

[4] P. Kairouz et al., “Advances and open problems in federated learning,” *Foundations and Trends in Machine Learning*, vol. 14, no. 1 - 2, pp. 1 - 210, 2021.

[5] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, “SCAFFOLD: Stochastic controlled averaging for federated learning,” *in Proc. of International Conference on Machine Learning (ICML)*, pp. 5132 - 5143, July 2020.

[6] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A.

Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” *in Proc. of Machine Learning and Systems (MLSys)*, 2020.
[7] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, “Federated learning with matched averaging,” *in Proc. of International Conference on Learning Representations (ICLR)*, Apr. 2020.
[8] Q. Li, B. He, and D. Song, “Model-contrastive federated learning,” *in Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10713 - 10722, June 2021.
[9] G. Lee, M. Jeong, Y. Shin, S. Bae, and S.-Y. Yun, “Preservation of the Global Knowledge by Not-True Distillation in Federated Learning,” *in Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
[10] Y. Shi, J. Liang, W. Zhang, V. Y. F. Tan, and S. Bai, “Towards Understanding and Mitigating Dimensional Collapse in Heterogeneous Federated Learning,” *in International Conference on Learning Representations (ICLR)*, 2023.
[11] L. Qu, Y. Zhou, P. P. Liang, Y. Xia, F. Wang, E. Adeli, L. Fei-Fei, and D. Rubin, “Rethinking Architecture Design for Tackling Data Heterogeneity in Federated Learning,” *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10061 - 10071, 2022.
[12] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” *in Proc. of International Conference on Learning Representations (ICLR)*, Apr. 2014.
[13] I. Higgins et al., “β-VAE: Learning basic visual concepts with a constrained variational framework,” *in Proc. of International Conference on Learning Representations (ICLR)*, Apr. 2017.
[14] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *in Proc. of International Conference on Learning Representations (ICLR)*, May 2021.
[15] A. Augenstein et al., “Generative models for effective ML on private, decentralized datasets,” *in Proc. of International Conference on Learning Representations (ICLR)*, Apr. 2020.
[16] Z. Zhu, J. Hong, and J. Zhou, “Data-free knowledge distillation for heterogeneous federated learning,” *in Proc. of International Conference on Machine Learning (ICML)*, pp. 12878 - 12889, July 2021.
[17] M. Polato, “Federated variational autoencoder for collaborative filtering,” *in Proc. of IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1 - 8, July 2021.
[18] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, “Ensemble distillation for robust model fusion in federated learning,” *in Proc. of Neural Information Processing Systems (NeurIPS)*, pp. 2351 - 2363, Dec. 2020.

저 자 소 개

박지우(준회원)

2025년 순천대학교 정보통신공학과 학사 졸업
2025년 순천대학교 정보통신공학과 석사 재학

<주관심분야 : 머신러닝, 딥러닝, 데이터분석>

강태욱(준회원)

2025년 순천대학교 정보통신공학과 학사 졸업
2025년 순천대학교 정보통신공학과 석사 재학

<주관심분야 : 머신러닝, 딥러닝, 데이터분석>

박철영(정회원)

2010년 순천대학교 정보통신공학과 학사 졸업.
2012년 순천대학교 정보통신공학과 석사 졸업.
2017년 순천대학교 정보통신공학과 박사 졸업.
2025년 순천대학교 인공지능공학부 조교수

<주관심분야 : 인공지능 응용시스템, 딥러닝, 머신러닝, 클라우드 컴퓨팅>

신창선(종신회원)

1996년 우석대학교 전산학과 학사 졸업.
1999년 한양대학교 컴퓨터 교육학과 석사 졸업.
2004년 원광대학교 컴퓨터공학과 박사 졸업.
2005년~현재 순천대학교 인공지능공학부 교수

<주관심분야 : 머신러닝, 분산시스템>