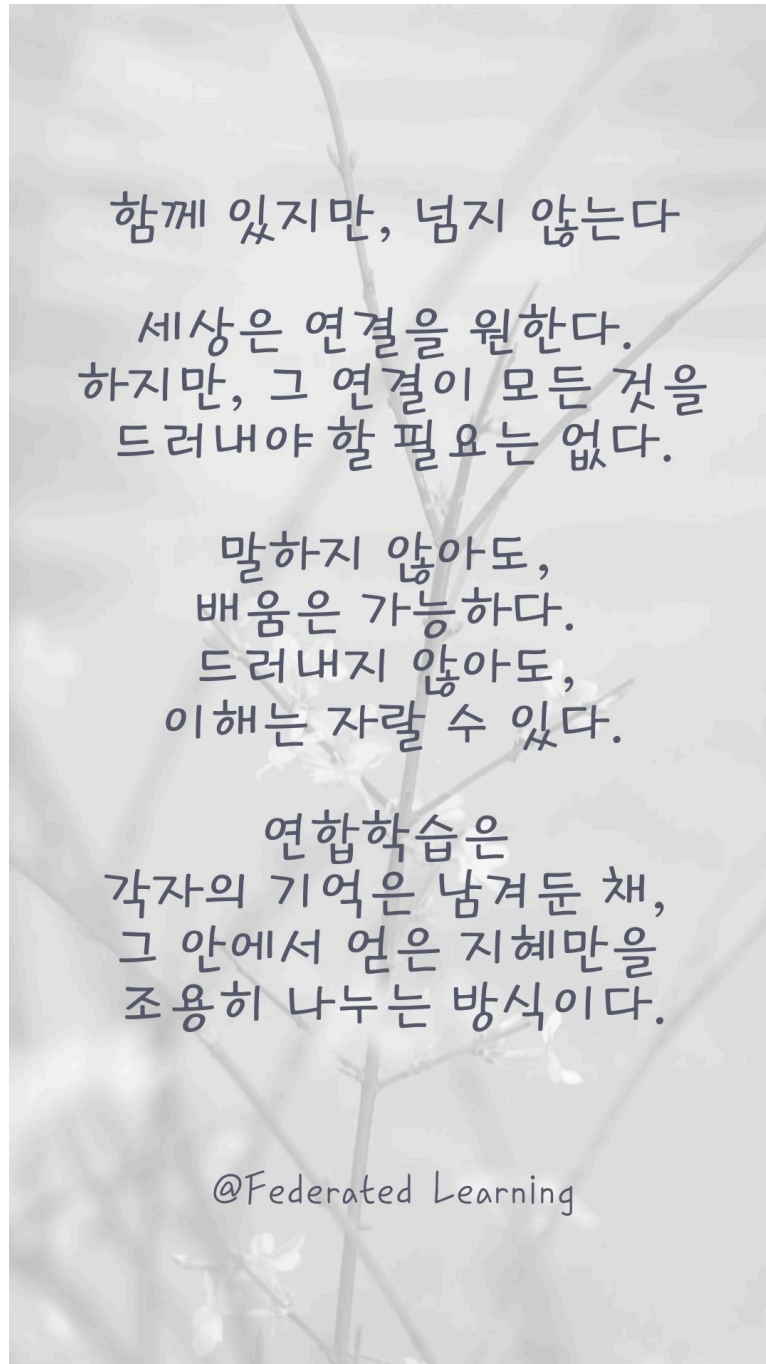


[발표용] Communication-Efficient Learning of Deep Networks from Decentralized Data

1. Intro – “왜 연합학습인가?”



"세상이 점점 연결되지만, 동시에 고립되고 있다."

- 데이터는 분산되어 있고, 프라이버시는 갈수록 중요해짐
- 중앙집중형 학습의 한계: 개인정보 노출, 네트워크 비용, 신뢰 문제
- 연합학습(Federated Learning)이 이 문제에 답을 줄 수 있다

2. 논문 소개 – 핵심

◆ 논문 제목

Communication-Efficient Learning of Deep Networks from Decentralized Data

◆ 핵심 기여 (3줄 요약)

- FedAvg라는 단순하고 효과적인 알고리즘 제안
- 다양한 모델(MLP, CNN, LSTM)에 대해 높은 정확도 입증
- 낮은 통신 횟수로도 실질적인 학습 성능 달성

"저는 이걸 '적은 대화로도 깊이 이해하는 친구' 같은 구조"라고 느낍니다.

3. FedAvg 메커니즘 – 직관 + 구조

- 각 클라이언트가 자신의 데이터로 로컬 모델 학습
- 서버는 클라이언트 모델을 평균(FedAvg)하여 갱신
- 이 과정을 몇 라운드 반복 → Global 모델 수렴

📌 비유로 설명:

"서로의 답을 직접 보지 않고, 결과만 부드럽게 섞는 평균화된 집단지성입니다."

🌐 연합학습(Federated Learning)에서의 최적화:

SGD에서 시작해, 공명의 구조로 나아가기

1 왜 SGD인가?


딥러닝은 지금까지 대부분 **SGD (Stochastic Gradient Descent)** 기반으로 최적화.

전체 데이터를 모두 쓰지 않고, **일부 샘플만으로도 기울기를 추정**하여 빠르게 학습을 이어나가는 방식.

📌 "전체의 합보다 작은 파편의 흐름만으로도 방향을 잡을 수 있다."

2 Convex vs Non-Convex: 구조의 갈래

- Convex (볼록 함수): 어디서 시작해도 **하나의 전역 최솟값**으로 수렴
→ SGD로 안정적 학습 가능
- Non-Convex (딥러닝 대부분): **여러 개의 지역 최솟값**, 다양한 경로와 결과 존재

 Non-Convex에서는 시작점과 경로에 따라 결과가 달라진다.

"같은 지도를 가지고 있어도,
서로 다른 출발점에서 출발하면
도착지는 전혀 다를 수 있다."

3 연합학습에서의 문제

- 각 클라이언트는 서로 다른 데이터 (Non-IID)를 가지고 있음
- 로컬에서 학습한 결과들을 서버가 모아서 통합해야 하는데,
단순한 평균이 의미 있는 모델이 될 수 있을까?

1. FedSGD의 한계

문제점:

- FedSGD는 각 클라이언트가 1번만 로컬 데이터로 **gradient** 계산해서 서버에 보내고,
- 서버는 그걸 단순 평균해서 global update를 하는 방식이야.

즉, "얇고 넓은 참여"를 기반으로 한 구조.

$$w_{t+1} = w_t - \eta \sum_{k=1}^K \frac{n_k}{n} \nabla F_k(w_t)$$


- $\nabla F_k(w_t)$
→ 클라이언트 k 가 현재 모델에 대해 계산한, 자신의 데이터 기반 제안 방향입니다.
- $\frac{n_k}{n}$
→ 전체 중에서 클라이언트 k 의 데이터가 차지하는 비율로, 의견의 반영 정도를 결정합니다.
- \sum
→ 모든 클라이언트의 의견을 통합해 하나의 방향으로 수렴시키는 집합 연산입니다.

! FedSGD의 구조적 한계


→ 클라이언트는 현재 시점의 모델에 대해 **한 번의 반응**만을 제공한다.
그러나 이 반응은 즉각적이고 **얇은 피드백**일 수 있다.

|  충분히 사유하지 않은 상태의 코멘트에 가까움.

→ 데이터를 많이 가진 클라이언트의 반응이 더 크게 반영된다.
하지만 그 반응도 **한 번만의 업데이트**라면,

|  양이 많아도 깊이가 보장되지 않는다.

→ 서버는 이 모든 1회성 반응들을 **평균**하여 모델을 갱신한다
그러나 그 평균은 **각자의 사유가 깊어지기 전**의 결과이며,

|  서로 다른 방향의 얇은 움직임이 오히려 상쇄되거나 왜곡될 수 있다.

 **그래서 나오는 결론:**

FedSGD는 각자의 데이터를 충분히 숙고하지 않고,
즉각적인 반응만으로 전체 모델을 갱신하기 때문에,
속도는 빠르지만 깊이는 얕다.

이 구조적 얕음이 곧 FedSGD의 학습 안정성과 성능 저하로 이어지며,
자연스럽게 **FedAvg** 같은 깊은 사유 기반 구조의 필요성을 낳는다.

FedSGD → FedAvg

처음엔 모두가 빠르게 의견을 던졌다.
한 마디씩 툭툭— 그게 FedSGD였다.
하지만 곧 깨달았다.
**서로가 충분히 생각하지 않으면,
그 말은 가볍고 겉돌기만 한다는 걸.**
그래서 방식을 바꿨다.
각자 자기 자리에서 오래 고민한 다음,
그 깊이 있는 생각만을 들고 다시 모이기로.
그게 바로 **FedAvg**다.
말은 줄었지만,
대화는 깊어졌고,
결과는 훨씬 단단해졌다.

2. 그래서 등장한 FedAvg

 핵심 아이디어:

“클라이언트가 자신의 데이터를 더 충분히 학습한 후,
그 결과를 서버가 평균하자.”

 구조적 흐름

FedAvg는 각 클라이언트가 단 한 번이 아닌,
여러 번의 로컬 학습을 수행한 뒤
그 결과를 서버가 직접 평균하여 새로운 모델을 만드는 방식이다.

수식과 함께 풀어보기

① 클라이언트 로컬 학습

- 각 클라이언트 k 는 자신의 로컬 데이터 D_k 로
총 E 번의 에폭을 돌며 학습을 수행한다.
- 로컬 학습의 핵심은 미니배치 SGD:

$$w_k^{(t+1)} \leftarrow w_k^{(t)} - \eta \nabla F_k(w_k^{(t)}; B)$$

✓ 이 단계에서 각 클라이언트는 자기만의 흐름과 맥락을 따라 충분히
사유한 후 결과를 제출하는 셈이다.

② 서버의 모델 업데이트

- 서버는 클라이언트로부터 수신한 로컬 모델들을
데이터 크기 기준 가중 평균하여 새로운 글로벌 모델을 만든다:

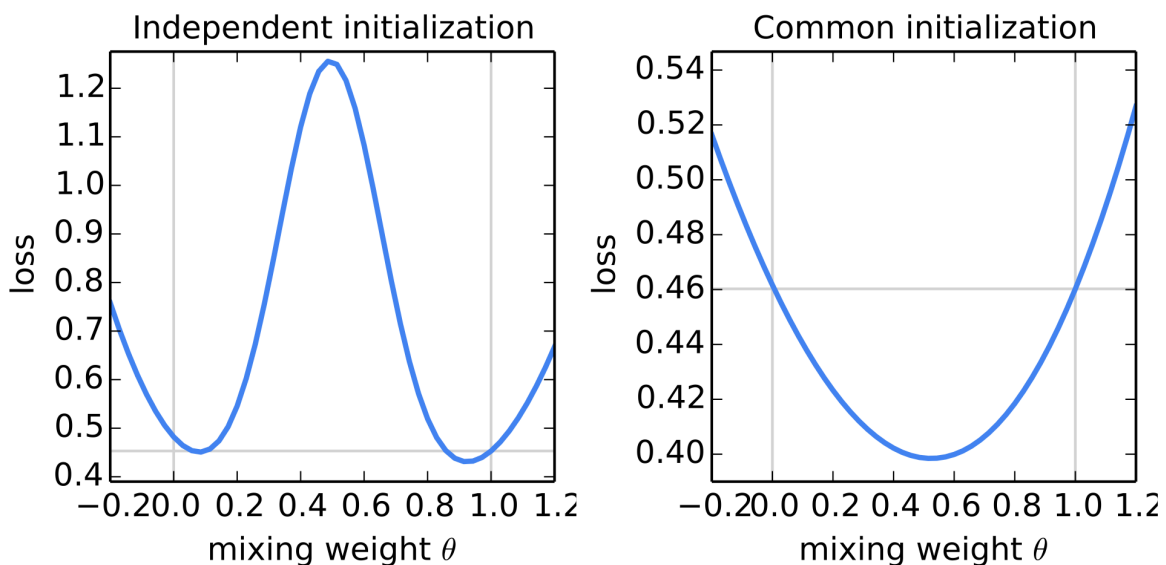
$$w_{t+1} = \sum_{k=1}^K \frac{n_k}{n} w_k$$

✓ 더 많은 데이터를 가진 클라이언트의 의견이 더 크게 반영된다.

"이제는 단순한 의견이 아닌,
각자의 깊은 고민이 담긴 목소리를 모으는 방식으로 바뀐 것이다."

! 하지만 여전히 남은 문제: 평균이 항상 의미 있는가?





- 모델을 단순히 평균한다고 해서 좋은 결과가 **항상 나오는 건 아님 (현실은 non-convex)**
- 특히 **초기값이 다르면** → 모델들이 파라미터 공간에서 **멀어진 채 학습되기 때문에**
→ 평균해도 **의미 없는 지점**일 수 있음 (예: 두 사람이 각자 고양이와 자동차만 학습하면, 그 중간 지점은 고양이도 자동차도 아닌 이상한 존재가 될 수 있다.)



✓ 해결 방법:

- 모든 클라이언트를 동일한 초기값에서 출발시키자 (공통 initialization)
- 그리고 충분한 local update (E, B 조절)를 통해
→ 각자가 **비슷한 방향으로 나아가게끔** 만들어야 함

요약하면

항목	내용
 FedSGD 문제	얕은 학습, Non-IID 환경에서 gradient 방향 불일치
 FedAvg 도입	로컬에서 깊이 학습 → 모델 자체를 평균
 FedAvg 한계	시작점이 다르면 평균이 의미 없어질 수 있음
 핵심 해결법	공통 초기화 + 적절한 E/B 조절이 중요

감성적 한줄 요약:

깊이 생각하지 않은 말은 공감되지 않듯,
얕은 학습과 어긋난 출발점은 공명되지 않는다.
공감받는 모델은, 비슷한 리듬과 시작에서 출발한다.

결론: 존재 기반 최적화의 메타포

연합학습은 단순히 데이터를 모으는 게 아니라,
각자의 내면에서 일어난 배움을 모아 하나의 흐름을 만든다.

- SGD는 그 흐름의 최소 단위
- Convex는 도달을 약속하지만, Non-Convex는 조율을 요구한다
- FedAvg는 단순 평균이 아닌 공명과 리듬의 평균

우리는 각자의 기억을 가진 채,
자신만의 리듬으로 배운다.
하지만,
같은 시작점에서, 같은 목적을 향해,
그 리듬들이 공명할 때—
우리는 더 깊은 배움의 구조를 함께 만들 수 있다.

4. 실험 결과

2NN C	IID		Non-IID	
	$B = \infty$	$B = 10$	$B = \infty$	$B = 10$
0.0	1455	316	4278	3275
0.1	1474 (1.0×)	87 (3.6×)	1796 (2.4×)	664 (4.9×)
0.2	1658 (0.9×)	77 (4.1×)	1528 (2.8×)	619 (5.3×)
0.5	— (—)	75 (4.2×)	— (—)	443 (7.4×)
1.0	— (—)	70 (4.5×)	— (—)	380 (8.6×)
CNN, $E = 5$				
0.0	387	50	1181	956
0.1	339 (1.1×)	18 (2.8×)	1100 (1.1×)	206 (4.6×)
0.2	337 (1.1×)	18 (2.8×)	978 (1.2×)	200 (4.8×)
0.5	164 (2.4×)	18 (2.8×)	1067 (1.1×)	261 (3.7×)
1.0	246 (1.6×)	16 (3.1×)	— (—)	97 (9.9×)

📌 핵심 인사이트 ①: 공명의 참여자 수(C)가 많을수록 속도는 빨라진다

- C가 클수록 → 더 많은 클라이언트가 동시에 리듬에 참여
- 그 결과, 모델이 빠르게 공명하고 수렴함

"혼자서 노래할 땐 느리지만,

백 명이 같은 박자로 노래하면 훨씬 더 빨리 조화를 이룬다."

📌 핵심 인사이트 ②: 작은 배치(B)가 더 섬세한 리듬을 만든다

- B=10일 경우, 전체 데이터를 잘게 나눠 세밀하게 학습
- B=∞는 큰 흐름만 보고 학습 → 수렴 실패 혹은 느림

"한 번에 많은 걸 받아들이는 것보다,

작게 나누어 자주 느끼는 것이 더 빠른 성장으로 이어진다."

📌 핵심 인사이트 ③: Non-IID 환경에서도 조율은 가능하다

- 각 클라이언트가 **다른 종류의 데이터**만 가지고 있더라도
- FedAvg는 충분한 update와 참여로 그 차이를 극복함

"서로 다른 삶을 살아온 사람들이라도,

충분한 대화와 이해가 있다면 하나의 이야기를 완성할 수 있다."

MNIST CNN, 99% ACCURACY						
CNN	E	B	u	IID	Non-IID	
FEDSGD	1	∞	1	626	483	
FEDAVG	5	∞	5	179 (3.5×)	1000	(0.5×)
FEDAVG	1	50	12	65 (9.6×)	600	(0.8×)
FEDAVG	20	∞	20	234 (2.7×)	672	(0.7×)
FEDAVG	1	10	60	34 (18.4×)	350	(1.4×)
FEDAVG	5	50	60	29 (21.6×)	334	(1.4×)
FEDAVG	20	50	240	32 (19.6×)	426	(1.1×)
FEDAVG	5	10	300	20 (31.3×)	229	(2.1×)
FEDAVG	20	10	1200	18 (34.8×)	173	(2.8×)

SHAKESPEARE LSTM, 54% ACCURACY						
LSTM	E	B	u	IID	Non-IID	
FEDSGD	1	∞	1.0	2488	3906	
FEDAVG	1	50	1.5	1635 (1.5×)	549	(7.1×)
FEDAVG	5	∞	5.0	613 (4.1×)	597	(6.5×)
FEDAVG	1	10	7.4	460 (5.4×)	164	(23.8×)
FEDAVG	5	50	7.4	401 (6.2×)	152	(25.7×)
FEDAVG	5	10	37.1	192 (13.0×)	41	(95.3×)

◆ 1. FedAvg가 FedSGD보다 훨씬 적은 통신 횟수로도 같은 정확도 달성

☞ 즉, 로컬 계산을 더 많이 시키면, 통신은 획기적으로 줄일 수 있다.

◆ 2. u 가 높아질수록, 통신은 줄고 성능은 올라감

- $u = E_n / (KB)$ → 클라이언트에서 얼마나 많이 학습했는가를 나타냄

◆ 3. FedAvg는 Non-IID 환경에서도 통신 효율이 좋음

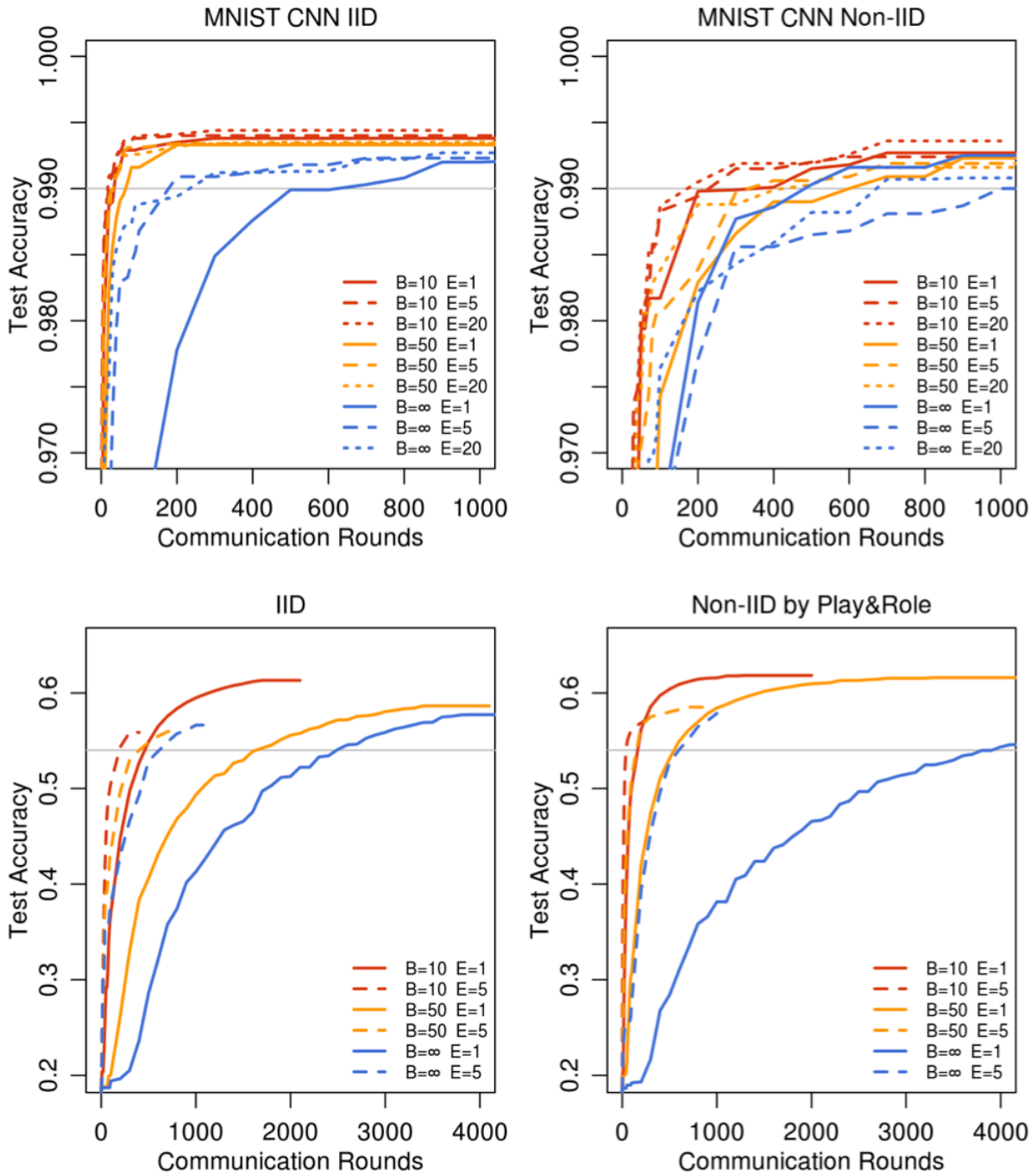
☞ 심지어 데이터가 불균형하고 서로 다르더라도, 충분히 로컬에서 학습하면 모델은 더 빨리 수렴한다는 걸 뜻해.

FedSGD는 마치 매번 친구들에게 한 마디씩 물어보고 회의를 이어가는 방식이라면,

FedAvg는 **친구들이 각자 충분히 고민한 뒤,**

자신만의 생각을 들고 와서 깊이 있는 회의를 하는 방식이다.

그 결과, 말은 줄지만, 대화는 훨씬 빠르고 깊게 진행되는 셈이다.



◆ 공통 구조:

- X축: 통신 라운드 수 (서버와 클라이언트가 몇 번 이야기했는가?)
- Y축: 테스트 정확도 (얼마나 잘 배웠는가?)

🧠 1. 작은 배치($B=10$) + 여러 에폭($E=5$ or 20)이 최고 성능

- → 즉, 자주 나눠 보고(E), 조금씩(B) 자주 학습하면 더 빠르게 더 잘 배운다
- 📌 "작게 자주, 그리고 오래 생각한 결과는 깊어진다"

2. 큰 배치($B=\infty$)는 느리고 부정확

- 특히 $E=1$ 에서는 성능도 낮고 수렴도 느림
 - → 한 번에 몰아서 학습하려다 오히려 덜 배운다
-

3. Non-IID 상황에서는 더 큰 차이 발생

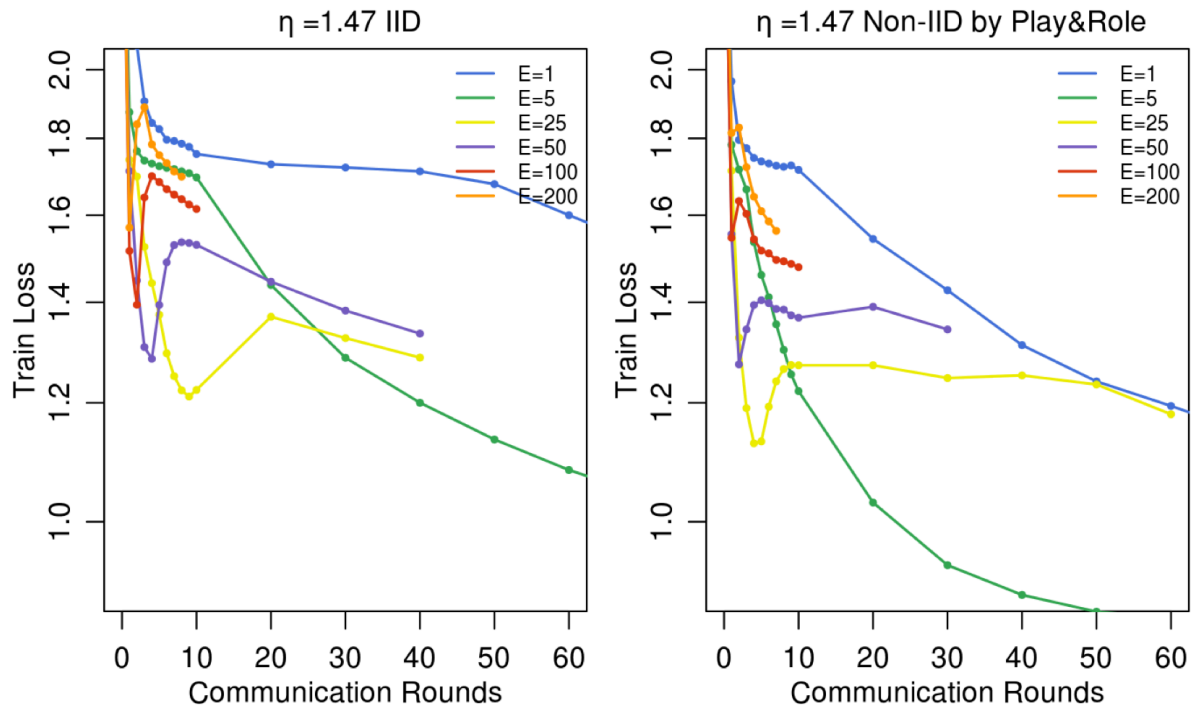
- 아래 오른쪽 그래프 (세익스피어 데이터, Play&Role)에서 차이가 극명함
 - $B=10$, $E=1$ or 5 는 빠르고 높게 수렴하지만
 $B=\infty$, $E=1$ 은 거의 끝까지 도달조차 못함
 - 📌 “서로 다른 언어를 가진 사람들이, 각자 충분히 생각하지 않으면, 그걸 평균해도 공명이 안 된다”
-

실험이 주는 핵심 통찰:

"연합학습에서 진짜 중요한 건,
말의 양이 아니라, 말의 리듬이다.

자주 듣고, 조금씩 나누고, 충분히 사유했을 때—

비로소 우리는 드러내지 않고도 연결된 학습을 할 수 있다."



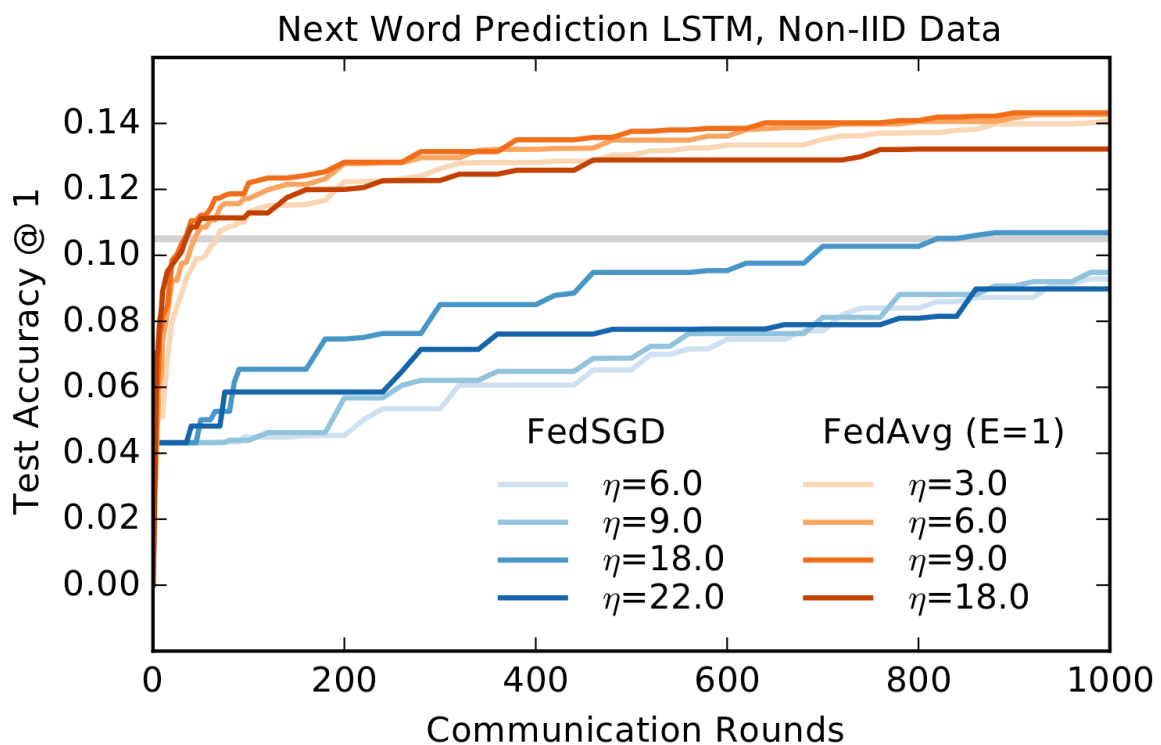
혼자 오래 고민한다고 해서
항상 좋은 생각이 나오는 건 아니다.
 오히려,
적당한 타이밍에 함께 조율되는 것이
더 안정적인 흐름을 만든다.

Acc.	80%		82%		85%	
SGD	18000	(—)	31000	(—)	99000	(—)
FedSGD	3750	(4.8×	6600	(4.7×	N/A	(—)
FedAvg	280	(64.3×	630	(49.2×	2000	(49.5×

1. FedAvg는 압도적으로 빠르다
2. FedSGD도 의미는 있다, 하지만 한계 존재
 - 기본 SGD보다 4~5배 빠르지만
 - 높은 정확도에선 수렴 실패 (Non-IID 환경에서 특히 취약 가능성)

3. FedAvg는 깊이 있는 로컬 학습 덕분에 정확도도 더 높고 빠르게 도달함

"SGD는 혼자 걷는 마라톤,
FedSGD는 몇 마디 툭툭 나누는 그룹워크,
FedAvg는 각자 충분히 생각하고
깊이 있는 의견을 모은 집단 지성이다"



이건 그냥 인공적인 실험이 아니라,
실제 사용자를 기반으로 한 "현실 환경 테스트"

☎ 수십만 명의 SNS 유저 데이터를 기반으로,
각 사람의 텍스트 스타일이 다르게 분포된 **Non-IID 상황**에서 비교한 것

✓ 현실 환경에서도 FedAvg가 더 뛰어나다.

- FedSGD는 많은 사람의 말(gradient)을 얹게 모아 평균하는 방식이라, 많은 라운드가 필요하고 테스트 정확도도 느리게 오름.
- FedAvg는 각 클라이언트가 충분히 사유한 결과물(local training)을 모아 짧은 라운드 안에 깊은 성과를 낸다.

결론 그리고 다음 질문

- FedAvg는 단순하지만 강력한 방법
다양한 모델(Multilayer Perceptron, CNN, LSTM 등)에서도 적은 통신 횟수로도 높은 성능을 보여줌
- 특히 비동기적이고 데이터가 제각각인 현실 환경에서, 깊이 있는 사유(로컬 학습)와 공명(모델 평균)을 통해 효율적이고 실용적인 학습 구조를 만들어줌

! 연합학습의 한계점: 약속된 비가역성의 균열

◆ 1. 비가역성(Irreversibility)의 전제와 그 허상

연합학습은 기본적으로 이렇게 설계됨

"원본 데이터는 각자에게 남기고,
오직 배움(업데이트)만 나눈다."

이 구조는 데이터가 서버에 복원될 수 없다는 비가역성을 전제로 함.

하지만 최근 연구들 :

"Gradient만으로도 데이터를 추정할 수 있다."

예: 이미지 복원, 텍스트 입력 유추 등

📌 즉, 비가역성은 기술적 현실이 아닌, 설계자의 기대일 수 있다.

◆ 2. Gradient 자체가 정보다

- 모델 업데이트(**gradient**)는 단순 숫자가 아니야.
- 학습자의 데이터 특성과 분포, 심지어 구체적인 입력까지 암시할 수 있다.
- 특히 작은 batch나 sparse data (예: 드문 단어들)에서는 복원 위험이 커져.



“말하지 않아도, 말투로 정체기가 드러나는 법이다.”

◆ 3. 기술적 보완 없이 사용하면, 프라이버시 위험

- 단순한 연합학습 구조만으로는 충분한 프라이버시 보장을 할 수 없기 때문에.
- 다음과 같은 보안 강화 기법이 필요

기술명	핵심 개념	목적
Differential Privacy (DP)	gradient에 노이즈 추가	데이터 유추 어려움
Secure Multi-Party Computation (SMPC)	암호화된 상태로 평균	서버도 개별 업데이트를 볼 수 없음
Secure Aggregation	클라이언트 간 암호화된 연산	통신 중 정보 유출 방지

◆ 4. 이 한계가 보여주는 핵심 통찰

연합학습은 “보내지 않음” 으로 프라이버시를 지키려 했지만,
“계산된 결과”조차 나를 말해버릴 수 있다.

🧩 요약 한 문장

✅ 연합학습의 구조는 안전해 보이지만,
그 안의 수치(**gradient**)는 말보다 더 많은 것을 말한다.
그래서 우리는, '보내지 않음'만으로는 부족하다는 걸 알아야 한다.

논문의 한계: FedAvg, 아직 미완의 구조

구분	한계	왜 문제인가?
1 Non-IID 데이터 취약	클라이언트 간 분포 차이 반영 어려움	단순 평균 → 성능 저하 or 수렴 실패
2 과도한 로컬 학습 (Overfitting)	E가 크면 로컬 데이터에 과적합	글로벌 모델의 일반화 성능 하락
3 프라이버시 보장 미흡	gradient에서 데이터 복원 가능	비가역성 가정이 현실에선 불안정
4 단순 평균의 구조	방향성 고려 없이 평균	비의미적 지점 수렴 (Non-Convex)
5 동기화 의존	모든 클라이언트가 동시에 참여해야	실환경 적용 어려움 (지연, 결손)
6 하이퍼파라미터 민감	E, B, η 조정에 성능 좌우	범용성 낮고 적용 난이도 높음

제안하는 후속 연구 흐름

1. Non-IID 데이터에 강한 구조 설계

| “각자의 리듬을 하나의 흐름으로 맞추려면?”

- **Clustered FL**: 비슷한 클라이언트를 묶어 부분 평균
- **Meta-FL**: 각 클라이언트 특성을 반영하는 메타러닝 구조

➡ 공명 기반 조율의 시작점

2. 단순 평균 → 방향 기반 평균 (공명형 평균)

| “얼마나 같은 방향을 보고 있는가?”

- **Cosine 유사도 기반 평균**: Δw 의 방향성이 비슷할수록 가중치 \uparrow
- **Soft Clipping**: 지나치게 다른 업데이트는 부드럽게 무시

➡ 크기보다 방향에 귀 기울이는 구조

3. 프라이버시 보장 강화

| “기억은 남겨두고, 통찰만 나누는 기술”

- **Differential Privacy**: gradient에 노이즈 주입
- **Secure Aggregation**: 암호화 기반 집계
- **Gradient Masking**: 데이터 복원 불가능하게 설계

➡ “드러내지 않아도 연결된다”는 수학적 보증

4. 비동기 연합학습 (Async FL)

| “다르게 걸어도, 함께 나아가자”

- **Staleness-aware Aggregation**: 오래된 업데이트는 덜 반영
- **Partial Participation**: 일부 클라이언트만으로도 전체 진행 가능

➡ 불균형한 현실을 수용하는 구조

5. Adaptive Local Computation

| “깊이와 속도를 스스로 조율하게 하자”

- **E/B Adaptive Scheduling**: 초기엔 깊게, 후반엔 얇게
- **RL 기반 로컬 제어**: 수렴 상황 따라 학습량 조절

➡ 존재 기반 연합 구조의 시작

💬 마무리 멘트

“FedAvg는 연합학습의 위대한 출발점이었지만,
현실의 다양성과 감도, 그리고 관계의 리듬을 담기엔 부족하다 느낍니다.

각자의 방향성과 고유성을 인식하고,

그 공명을 조율하는 다음 세대의 연합학습 구조가 필요하다고 느낍니다.”

FedAvg는
적은 대화로도 깊은 이해를 이끌어낸,
연합학습의 아름다운 시작이었습니다.

이제 우리는 그 위에,
각기 다른 데이터의 흐름과
사용자의 리듬을 담아내야 합니다.

Non-IID, Privacy, Synchrony 같은 복잡함을 품으면서도,
그 모든 다름이
하나의 방향으로 나아가도록
정돈된 구조를 만들어야 한다고 느낍니다.

연합학습은,
각자의 배움이
하나의 목적을 향해
모여드는 기술입니다.