2024 SAIL Seminar

Communication-Efficient Learning of Deep Networks from Decentralized Data

H. Brendan McMahan, Eider Moore, et al., arXiv 2016 (# of citation: 13496)

SAIL
Senseable Artificial Intelligence Laboratory

순천향대학교 미래융합기술학과

Senseable AI 연구실

석사과정 김병훈

Index
Contents

1 Introduction

1. Introduction

Background and Requirement

Backgrounds

- Many people use mobile phones
- Mobile phones have data that contains personal information
- Training models with this data maximises usability for users

Problems

- Traditional centralised processing can expose privacy risks
- Centralisation of data creates bottlenecks

Requirements

- Couldn't we advance the model without transferring data?

1. Introduction

Contribution

Solution

- Introducing federated learning, a technique for training shared models without having to store rich data centrally.

Summary of how it works

- Combine the server performing the model average with each client performing the local SGD

Contribution

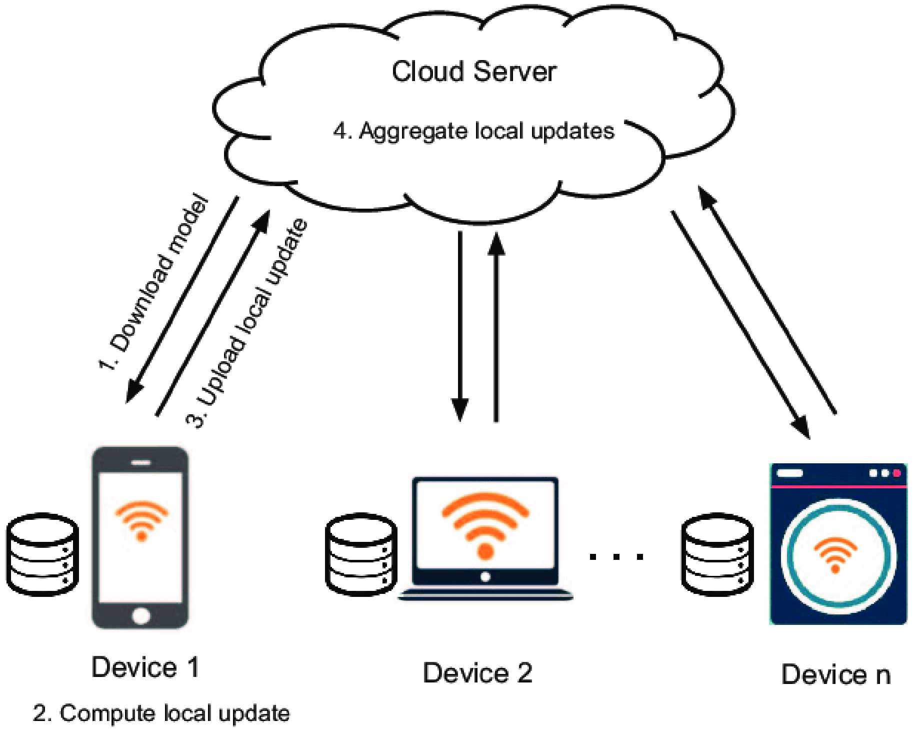
- Decentralising to solve bottlenecks
- Simple and practical algorithm using SGD and model averaging
- Extensive empirical evaluation

1. Introduction

Basic Concepts

Defining terms

- Non-IID: means that the data held by each distributed node is of a very different character and an imbalance of data exists.
- Data imbalance: Due to variations in mobile phone usage among users, there is an imbalance in the amount of data collected.



1. Introduction

Related works

Paper	Year	Methods	Limitation
Distributed training strategies for the structured perceptron [28]	2010	Averaging local training models	Data imbalance Not considering non-IID
Parallel training of deep neural networks with natural gradient and parameter averaging [31]	2015		
Information-theoretic lower bounds for distributed statistical estimation with communication constraints [45]	2013	Make distributed data communication more efficient	Data imbalance Not considering non-IID Few clients
Communication efficient distributed optimization using an approximate newton-type method [34]	2013		
Trading computation for communication: Distributed stochastic dual coordinate ascent [40]	2013		
Adding vs. averaging in distributed primal-dual optimization [27]	2015		
Communication-efficient distributed optimization of self-concordant empirical loss [43]	2015		
Communication-efficient algorithms for statistical optimization [44]	2012	Global model averaging	Not considering non-IID Performance issues
Communication complexity of distributed convex learning and optimization [3]	2015		
Parallelized stochastic gradient descent [46]	2010		

2 The FedAvg. Algorithm

2. The FedAvg. Algorithm

Baseline

Stochastic Gradient Descent (SGD) performs the gradient calculation for one batch of clients (randomly selected) in one round.

- Use Large-batch because it doesn't cost much for a large number of clients

η : learning rate
k : number of clients
n : number of data samples
w_t: current model weight

$$w_{t+1} \leftarrow w_t - \eta \Sigma_{k=1}^{K} \frac{n_k}{n} g_k \qquad \text{where} \qquad \Sigma_{k=1}^{K} \frac{n_k}{n} g_k = \nabla f(w_t)$$

$$w_{t+1}^{k} \leftarrow w_t - \eta g_k \qquad \text{and then,} \qquad w_{t+1} \leftarrow \Sigma_{k=1}^{K} \frac{n_k}{n} w_{t+1}^{k}$$

2. The FedAvg. Algorithm

FedAvg algorithm's pseudo code

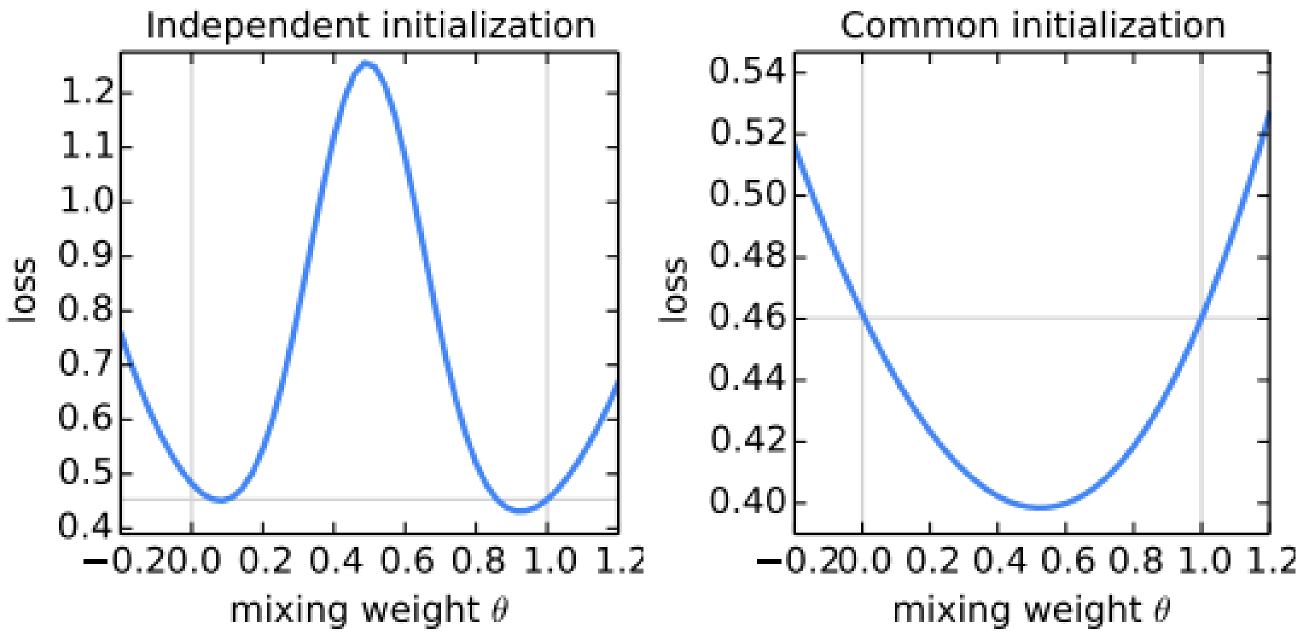
Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
  initialize w_0
  for each round $t = 1, 2, \dots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate(k, w)**:** *// Run on client k*
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return w to server

2. The FedAvg. Algorithm

Demonstrate effectiveness



Learning about small datasets

3 Experimental Results

3. Experimental Results

Datasets and Experimental design

Datasets

- MNIST (2NN and CNN)
- CIFAR-10 (2NN and CNN)
- The Complete Works of William Shakespeare (LSTM)

Experimental design

1. Impact of client participation rate C
2. Per-client Computation amount (Batch size and Epoch)
3. Evaluate FedSGD / FedAvg Algorithms Performance

3. Experimental Results

Impact of client participation rate C

MNIST

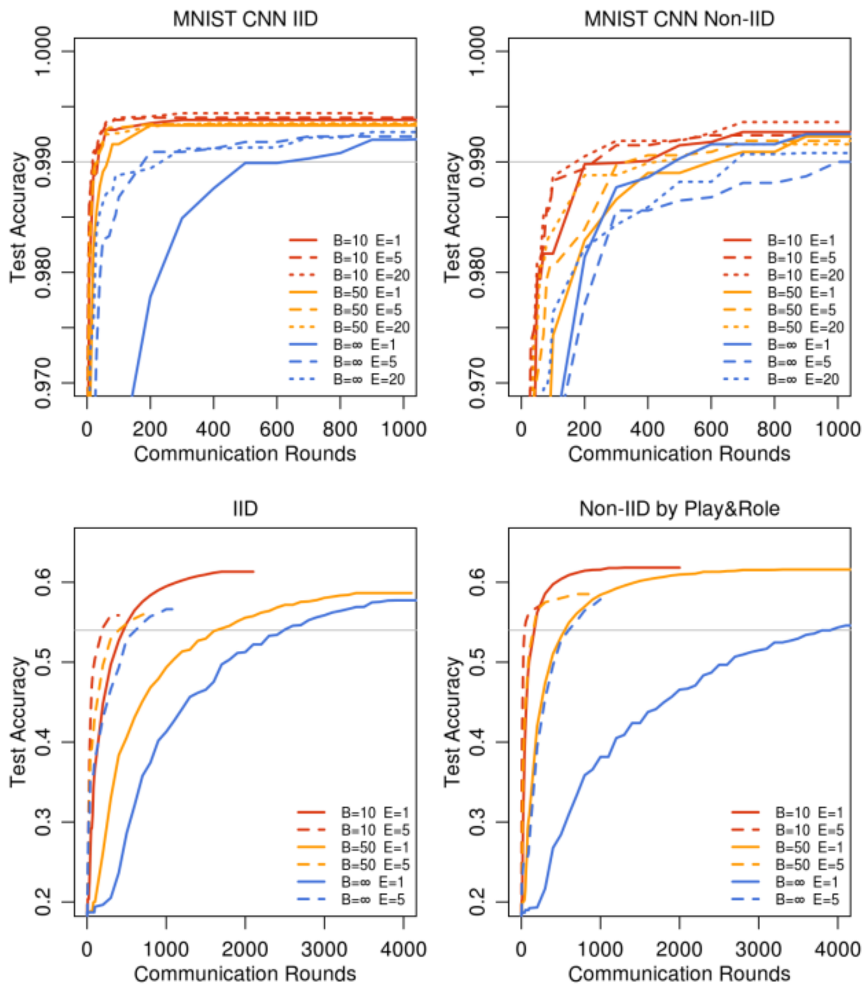
2NN	IID		NON-IID	
C	$B = \infty$	$B = 10$	$B = \infty$	$B = 10$
0.0	1455	316	4278	3275
0.1	1474 (1.0×)	87 (3.6×)	1796 (2.4×)	664 (4.9×)
0.2	1658 (0.9×)	77 (4.1×)	1528 (2.8×)	619 (5.3×)
0.5	— (—)	75 (4.2×)	— (—)	443 (7.4×)
1.0	— (—)	70 (4.5×)	— (—)	380 (8.6×)
CNN, $E = 5$				
0.0	387	50	1181	956
0.1	339 (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×)

Set up the rest of the experiment with C=0.1

3. Experimental Results

Per-client calculations(Batch size and Epoch)

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×)



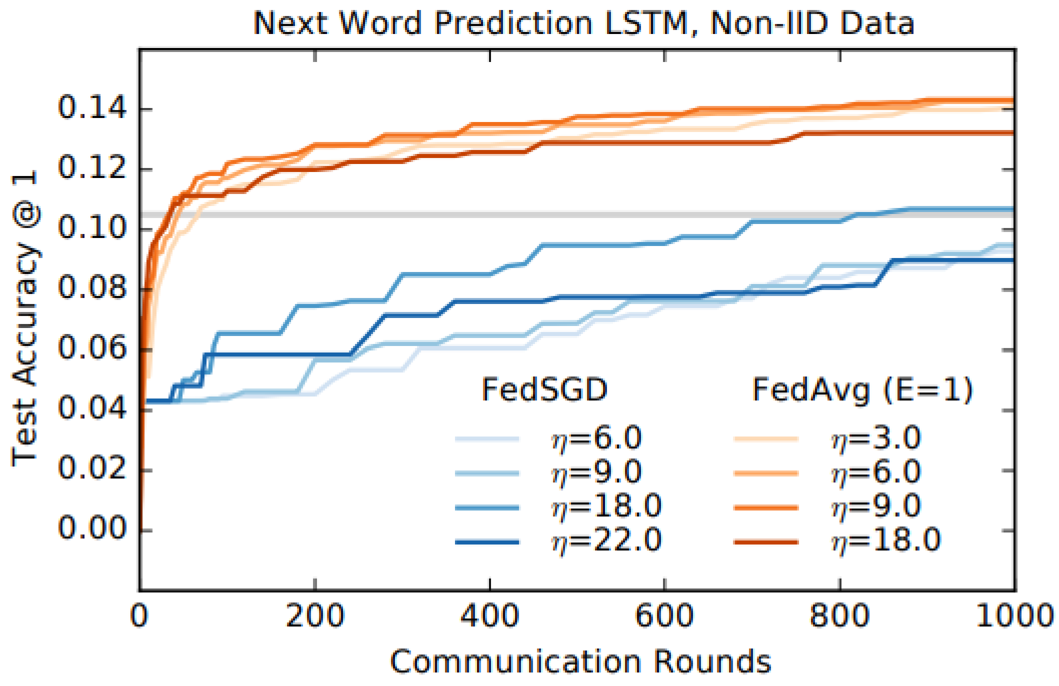
3. Experimental Results

Evaluate FedSGD / FedAvg Algorithms Performance

CIFAR-10

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×)

The Complete Works of William Shakespeare



4 Conclusion and Future Work

5. Conclusion and future work

Conclusion

Conclusion

1. 적은 수의 통신으로 고성능 모델을 얻을 수 있음
2. 실용적인 알고리즘

Future work

1. 개인정보 보호를
2. differential privacy나 secure multi-party computation등의 기술들을 적용

5. Conclusion and future work

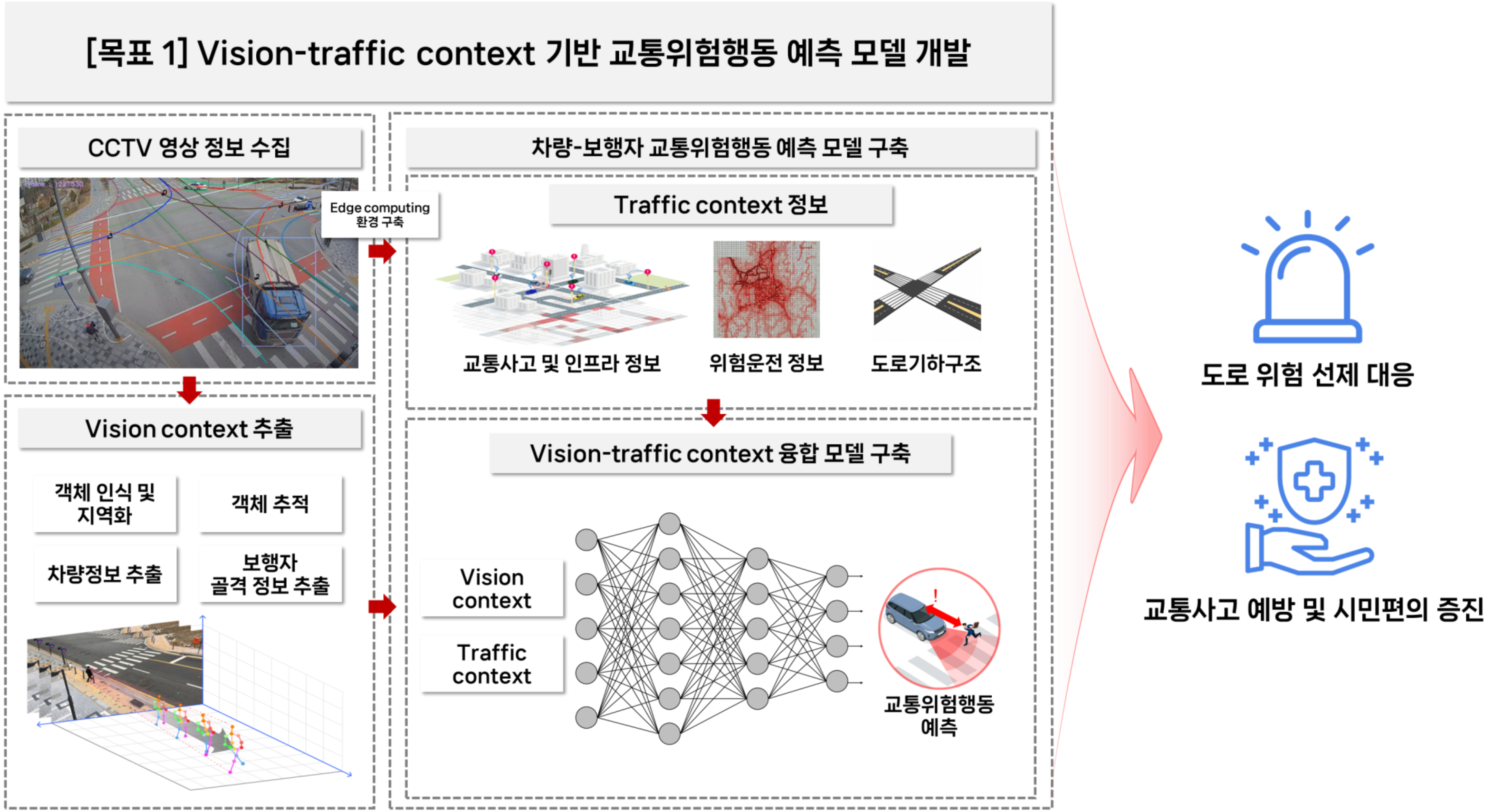
How to apply?

2024년도 석사과정생연구장려금지원사업
신규과제 연구계획서

과제명	국문	차세대 지능형 교통안전시스템(C-ITS)을 위한 영상기반 도로위험행동 예측모델의 맞춤형 준지도 연합 학습
	영문	Customised semi-supervised federated learning of video-based traffic risk behaviour prediction models for Cooperative-Intelligent Transport Systems(C-ITS)

5. Conclusion and future work

How to apply?



5. Conclusion and future work

How to apply?