https://doi.org/10.7840/kics.2023.48.2.266

업로딩 지역에 따른 연합 학습 성능 분석 및 최적화

전 성 훈◆, 김 동 인°

Analysis and Optimization about Federated Learning in Uploading Zone

Seong Hoon Jeon◆, Dong In Kim°

요 약

본 논문에선 자유롭게 분포된 디바이스가 연합 학습에 참여하기 위한 업로딩 지역을 제안한다. 업로딩 지역은 디바이스의 위치에 따른 경로 손실을 고려한 양자화 레벨에 따라 내곽과 외곽, 두 영역으로 나눠진다. 내곽 영역에 위치한 디바이스는 기지국과 가깝기 때문에 충분한 비트를 양자화 과정에 활용하지만, 외곽 영역에 위치한 디바이스는 기지국과 멀리 떨어져 있어 상대적으로 적은 비트를 양자화 과정에 활용한다. 따라서 이러한 양자화 이질성 환경을 고려하여 전역 학습 성능을 최대화하도록 로컬 학습을 최적화하였다. 아울러, 학습 성능을 최대로 하는 최적의 업로딩 지역의 크기를 찾기 위한 최적화 문제를 제안한다. 위의 결과는 다양한 개인 데이터를 가진 많은 수의 IoT 센서로 구성된 AI 인지 IoT 네트워크에 활용될 수 있다.

키워드 : 연합 학습, 양자화 이질성, 업로딩 지역, 동적 로컬 학습, IoT

Key Words : Federated Learning, Heterogeneous Quantization, Uploading Zone, Adaptive Local Update, Internet of Things (IoT)

ABSTRACT

In this paper, we propose an *Uploading Zone* where devices can participate in federated learning, given they are distributed randomly. Here, the uploading zone is divided into the two regions, inner and outer regions, according to the quantization levels considering the path loss depending on device location. Some devices located in inner region utilize a sufficient number of bits for quantization because they are close to Base Station (BS), but the devices located in outer region use a fewer bits for quantization because they are far from BS. Hence we optimize the local update under this heterogeneous quantization condition, so as to maximize the global learning performance. To this end, we formulate an optimization problem to determine the optimum size of uploading zone that leads to the maximum learning performance. This finding can serve as an appropriate framework for AI-native IoT Networks which consist of a huge number of IoT sensors with diverse private data.

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT명품인재양성 사업의 연구결과로 수행되었음 (IITP-2022-2020-0-01821)
◆ First Author : Department of Electrical and Computer Engineering, Sungkyunkwan University, shjeon96@g,skku.edu, 학생회원
° Corresponding Author : Department of Electrical and Computer Engineering, Sungkyunkwan University, dongin@skku.edu, 종신회원
논문번호 : 202209-230-B-RE.R1, Received September 30, 2022; Revised December 5, 2022; Accepted December 11, 2022

Ⅰ. 서 론

최근 디바이스에 탑재된 연산 능력의 비약적인 향상으로 인하여 이를 활용하는 인공지능 (Artificial Intelligence) 분야에 관한 연구가 활발히 진행되고 있다. 기존 인공지능은 학습에 필요한 매우 많은 수의 데이터를 모두 중앙 서버가 소유하고 있어야 하며 학습 과정 또한 중앙 서버에서만 이루어진다. 하지만 다양한 IoT (Internet of Things) 센서로부터 디바이스가 데이터를 수집하는 미래 IoT 네트워크의 경우, 학습에 필요한 데이터를 모두 디바이스가 중앙 서버로 전송하는 것은 통신 측면에서 매우 비효율적이며, 최근 민감하게 받아들여지고 있는 데이터 프라이버시 문제를 초래한다는 단점이 존재한다. 이를 해결하는 방법으로 최근 연합 학습(Federated Learning)이 제안되었다. 연합 학습은 개인 정보를 포함하는 학습데이터 대신 데이터와 무관한 상대적으로 규모가 작은 학습 모델을 공유하기 때문에 이러한 문제들을 효과적으로 해결할 수 있다는 장점이 있다[1].

하지만 최근 이미지 분류와 같이 AI를 활용하여 해결하고자 하는 학습 문제가 복잡해짐에 따라 원활한 학습을 위하여 큰 규모의 학습 모델이 필요하게 되었다. 따라서 학습 모델을 공유하는 연합 학습에서 역시, 모델을 공유하는 과정에서 존재하는 통신 병목현상을 효율적으로 접근하는 연구가 FL에 매우 중요한 연구 과제로 여겨지고 있으며, 이러한 연구들은 크게 2가지 측면으로 나눌 수 있다.

우선 학습 측면에서, 로컬 학습 개념을 활용하면 기존 연구에 비하여 매우 빠르게 학습이 수렴하는 것을 여러 연구를 통해 확인할 수 있다[2,3,7-9]. 이는 기존 FL과 달리 각 디바이스가 보유한 학습 데이터를 활용하여 로컬에서 여러 번 학습하여 생성된 학습 모델을 전송하는 방법으로, 학습 로스를 빠르게 감소시켜 학습이 수렴하기까지 필요한 통신 횟수 자체를 감소시킨다는 점에서 효과적이다. 하지만 J. Wang et al.의 연구[3]는 고정된 로컬 학습 횟수와 수렴된 학습 성능 사이에 Trade-Off가 존재한다는 것을 확인하여, 일정 시간 간격마다 로컬 학습 횟수를 최적화하는 동적 로컬 학습 최적화 방법을 제안하였다.

또한 통신 측면에서, 전송하는 학습 모델의 모델 파라미터인 gradient를 기존 통신 기법의 압축 센싱과 양자화를 활용하여 실제로 전송하는 학습 모델의 bit 수를 감소시켜 통신 시간을 조절하는 다양한 연구들이 존재한다[4-7]. D. Alistarh et al.의 연구[4]는 학습 모델의 layer들을 각각 2, 4, 8 bit로 통계적으로 양자화하며, 양자화로 인한 오류를 고려한 학습 성능을 분석하였다. 또한 P. Liu et al.의 연구[5]는 통계적 양자화 기법을 적용하면서 동시에 각 디바이스에게 최적의 주파수를 할당하는 방법을 연구하였다. 해당 연구에선 이러한 주파수 할당과 양자화 비트를 한 번에 최적화하는 대신 각각을 고정하여 최적화한 후 이를 종합하는 방법으로 학습 시간을 최소화하도록 양자화 비트와 주파수 할당 최적화 방법을 제안하였다. 그리고 D. Jhunjhunwala et al.의 연구[6]에선 통계적 양자화를 활용하지만 일정 라운드마다 사용하는 양자화 비트 수를 동적으로 최적화하는 방법을 연구하였다. 학습 초반에는 상대적으로 적은 비트 수를 활용하는 부정확한 양자화를 사용하고 이후 학습이 진행되면서 더 많은 비트 수를 양자화에 활용하여 gradient를 정확하게 표현하여 학습이 성공적으로 수렴할 수 있도록 하였다. 이후 다른 연구[7], [8]에선 block floating point 양자화 기법을 활용하여 gradient를 양자화하고, 모든 디바이스들이 각자 2가지 양자화 레벨 중 하나를 선택하는 양자화 이질성 환경에서 글로벌 모델을 업데이트하는 방법을 제안하였다. 또한 양자화 방법 대신 압축 센싱을 적용하여 학습 모델을 전송하는 M. K. Nori et al.의 연구[9]에선 앞선 동적 로컬 학습을 활용하면서 매 라운드마다 압축 정도와 로컬 학습 횟수를 최적화하여 로컬 학습 횟수만을 최적화하는 이전 연구[3]와 비교하여 압축 센싱을 추가로 최적화하는 것을 통하여 학습이 훨씬 빠르게 수렴하는 것을 확인하였다.

하지만 로컬 학습을 활용하는 이러한 연구[2,3,7,8]들은 동일한 로컬 학습을 활용한다는 한계가 존재하며, 동적 로컬 학습을 고려한 연구[3], [8]들 역시 모든 디바이스가 학습에 참여하며 동일한 전송률을 가진 환경을 가정하여 연구를 진행하였다. 또한 다른 연구들[4], [5], [6]은 학습 시간을 효과적으로 감소시킬 수 있는 로컬 학습을 고려하지 않고 진행되었다. 즉, 이러한 선행 연구들은 연합 학습에서 각 디바이스가 가진 하드웨어적, 통신적인 시스템 이질성을 고려하지 않은 한계가 존재한다. 따라서 본 연구에선 디바이스들이 분포된 환경에서 업로딩 지역[10]을 다시 정의하고, 디바이스가 겪는 통신 채널의 차이를 고려하여 양자화하는, 양자화 이질성이 존재하는 환경[11]을 가정한다. 이후 학습을 빠르게 수렴시키기 위한 로컬 학습 개념을 적용하기 위하여 양자화 이질성을 고려한 동적 로컬 학습 최적화 방법을 제안한다. 끝으로 업로딩 지역의 크기를 2가지로 나누어 학습 성능을 비교 및 분석하고, 학습 성능을 최대화하도록 업로딩 지역의 크기를 최적화 문제로 가져가 최종적으로 최적의 업로딩 지

역의 크기를 제안한다.

Ⅱ. 본 론

2.1 업로딩 지역과 디바이스 분포

N_{total}개의 디바이스들은 반지름으로 r_{end}를 갖는 임의의 원형 지역인 ***디바이스 지역*** 내부에 자유롭게 분포된다. 디바이스의 분포는 uniform distribution을 만족하며, 디바이스 j가 기지국(base station)으로부터 떨어진 거리를 d_j라고 정의되며, 이때 디바이스 j는 $f_D(d_j)=\frac{2d_j}{(r_{end})^2}$의 확률 분포를 따른다.

또한 디바이스가 학습에 참여하기 위한 물리적인 지역을 ***업로딩 지역***으로 정의하고, 업로딩 지역의 크기는 중앙 BS으로부터 거리 r_{zone}으로 정의한다. 따라서 디바이스 j가 학습에 참여하기 위한 조건은 업로딩 지역 내부에 위치해야 하며 디바이스의 학습 참여 확률과 이로 인한 통계적으로 학습에 참여하는 평균 디바이스 수는 각각 식 1, 2와 같이 정리된다.

$$p_{zone}=F_D(d_j\le r_{zone})=\left(\frac{r_{zone}}{r_{end}}\right)^2 \tag{1}$$

$$|\mathbb{J}|=\mathbb{E}[N_{zone}]=N_{total}\times p_{zone} \tag{2}$$

업로딩 지역의 크기로 인한 학습 성능을 분석하기 위하여 서로 다른 크기를 갖는 2개의 업로딩 지역, Small Zone과 Large Zone을 가정한다. 앞서 가정한 디바이스의 분포를 활용하면 두 업로딩 지역에서 학습에 참여하는 디바이스의 통계적 비율은 $\frac{\mathbb{E}[N_{small}]}{\mathbb{E}[N_{large}]}=\left(\frac{r_{small}}{r_{large}}\right)^2$으로 표현할 수 있다.

2.2 양자화 규칙

디바이스가 학습 모델을 전송하는 과정에서 모든 디바이스가 동일한 하드웨어 특징을 갖는다고 가정하면, 양자화 과정에서 중요한 요소는 디바이스의 위치로 인해 생기는 전송 경로 손실(Path Loss)이다. 따라서 PL를 고려한 디바이스 j의 데이터 전송률은 식 3과 같이 표현되며, 이때 P_T는 전송 전력을, B는 주파수대역을, N_0는 AWGN 채널의 주파수 power spectral density를, h_j는 채널 이득을, P_j는 BS이 디바이스 j로부터 받은 수신 전력을 의미한다.

$$R_j=R(d_j)=B\log_2\left(1+\frac{P_T h_j^2 d_j^{-\alpha}}{BN_0}\right)\approx\frac{P_j}{N_0}\log_2 e \tag{3}$$

따라서 디바이스의 위치를 고려하여 업로딩 지역 내부를 그림 1과 같이 2개의 영역으로 나누었다. 업로딩 지역은 상대적으로 BS과 가까워 충분히 많은 비트로 양자화 가능한 내부 영역과, BS으로부터 멀리 떨어져 있어 상대적으로 적은 비트로 양자화하는 외부 영역으로 구성된다.

본 연구에선 내부 영역에 위치한 디바이스는 Float16으로 모델 파라미터를 양자화하여 전송하며, 외부 영역에 위치한 디바이스는 Float8을 사용한다. 또한 업로딩 지역 바깥에 위치한 디바이스는 학습 모델을 전송하기에 충분한 전송률을 갖고 있지 않다고 판단되어 학습에 참여하지 못한다. 이러한 디바이스들의 양자화 규칙을 위하여 학습 참여 기준 전송률 R_{join}과 양자화 기준 전송률 R_Q는 아래와 같다.

$$R_{join}=R(r_{zone}),R_Q=2R_{join} \tag{4}$$

이때, 디바이스가 활용하는 양자화 구조는 그림 2와 같으며, 각 양자화 방법은 다음과 같은 성질을 만족한다: (i) 모든 양자화 과정은 Unbiased Estimator 과정으로 $\mathbb{E}[Q(x,q_j)]=x$를 만족하며, (ii) 양자화 과정의 오류 분산 값은 $\mathbb{E}[\|Q(x,q_j)-x\|^2]\le c_j\|x\|^2$로

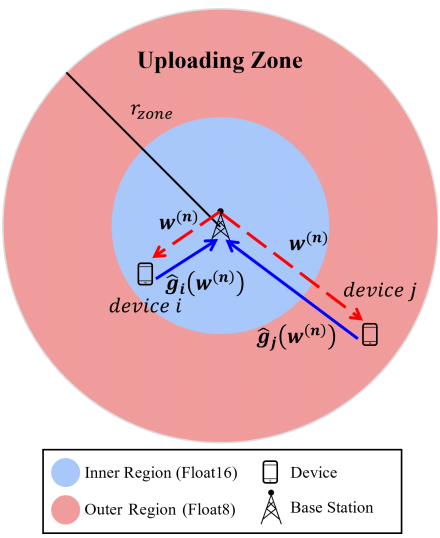


그림 1. 업로딩 지역 개요
Fig. 1. Illustration of Uploading Zone

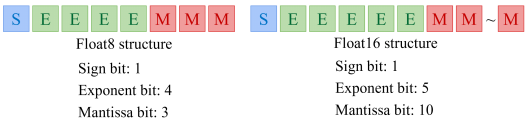


그림 2. Float8과 Float16 양자화 구조
Fig. 2. Structure of Float8 and Float16

상한값을 가진다. 이때, q_j는 $q_j \in \{8,16\}$을 만족하며, c_j는 양자화 방법에 따라 floor method에선 2^{-2M_j} 값을 가지며, round method에선 2^{-2M_j-2}의 값을 갖는다. 이때, M_j는 각 양자화 방법에서 가수 bit의 수를 의미한다.

따라서 각 디바이스들의 전송 시간은 앞서 정의한 식 3과 양자화 비트를 활용하여 아래와 같이 정의된다.

$$t_{UL_j}^{(n)} = \frac{q_j \times d}{R_j} \tag{5}$$

이때, q_j는 양자화 비트의 수로 $q_j \in \{8,16\}$를 만족하며 d는 모델 파라미터의 수를 의미한다. 각 디바이스가 본인의 전송률을 고려하여 Float8과 Float16 양자화 방법 중 선택하여 사용하기 때문에 모든 디바이스들의 상향 링크 전송 시간은 기준 전송률을 이용한 기준 전송 시간보다 항상 작은 것을 알 수 있다.

$$t_{UL_j}^{(n)} \le \frac{8 \times d}{R_{join}} = T_{zone} \tag{6}$$

2.3 로컬 학습 최적화

학습 로스를 빠르게 감소시키는 로컬 학습 기법을 활용하기 위해서는 제안하는 양자화 기법이 학습에 미치는 영향을 분석할 필요가 있다. 또한 기존 양자화를 활용한 연구[4], [6]과 달리 디바이스들이 2개의 양자화 level을 갖는 양자화 이질성 환경을 고려했기 때문에 양자화 이질성이 학습에 미치는 영향을 분석하면 식 7과 같다.

$$\begin{aligned}&\mathbb{E}\left[\|\hat{\boldsymbol{g}}(\boldsymbol{w}^{(n)}) - \nabla F(\boldsymbol{w}^{(n)})\|^2\right] \\ &= \mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{w}^{(n)}) - \nabla F(\boldsymbol{w}^{(n)})\|^2\right] \\ &\quad + \mathbb{E}\left[\|\hat{\boldsymbol{g}}(\boldsymbol{w}^{(n)}) - \boldsymbol{g}(\boldsymbol{w}^{(n)})\|^2\right]\end{aligned} \tag{7}$$

이때, $\boldsymbol{w}^{(n)}$은 n번째 라운드의 모델 파라미터를 의미하며, F는 loss function을, $\boldsymbol{g}$와 $\hat{\boldsymbol{g}}$는 각각 미니 배치를 활용한 SGD 기법의 gradient와 이를 양자화한 gradient 의미한다. 식 7은 앞서 양자화 과정의 Unbiased Estimator 특징으로 인하여 두 분산의 합으로 나눠지며, 첫 번째 항은 FGD와 SGD 과정의 차이로 인한 값을 의미한다. 이 항은 기존 SGD 연구에서 식 8과 같이 정의되어 활용된다.

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{w}^{(n)}) - \nabla F(\boldsymbol{w}^{(n)})\|^2\right] \\ &\le \beta \| \nabla F(\boldsymbol{w}^{(n)}) \|^2 + \sigma^2\end{aligned} \tag{8}$$

이때 β와 σ^2은 각각 미니 배치 ξ의 크기에 반비례하는 상수이다. 그리고 식 7의 2번째 항은 양자화 과정에서 생성되는 분산을 의미하며, 이때 사용하는 global gradient vector는 $\boldsymbol{g}(\boldsymbol{w}^{(n)}) =: \frac{1}{|\mathbb{J}_{zone}|} \sum_{j \in \mathbb{J}_{zone}} \boldsymbol{g_j}(\boldsymbol{w}^{(n)})$으로 정의된다. $\boldsymbol{g_j}(\boldsymbol{w}^{(n)})$은 디바이스 j의 local gradient를 의미한다. 따라서 식 7의 2번째 항은 다음과 같이 upper bound를 갖는다.

$$\begin{aligned}&\mathbb{E}\left[\|\hat{\boldsymbol{g}}(\boldsymbol{w}^{(n)}) - \boldsymbol{g}(\boldsymbol{w}^{(n)})\|^2\right] \\ &\le \frac{1}{|\mathbb{J}^{(n)}|} \sum_{j \in \mathbb{J}^{(n)}} (c_j) \\ &\times \frac{1}{|\mathbb{J}^{(n)}|} \sum_{j \in \mathbb{J}^{(n)}} \left(\mathbb{E}\left[\| \boldsymbol{g_j}(\boldsymbol{w}^{(n)}) \|^2\right]\right) \\ &\le \mathbb{E}_j[c_j] \times \mathbb{E}\left[\| \boldsymbol{g}(\boldsymbol{w}^{(n)}) \|^2\right]\end{aligned} \tag{9}$$

이때, 각 디바이스의 SGD 과정에서 분산을 의미하는 $\mathbb{E}\left[\| \boldsymbol{g_j}(\boldsymbol{w}^{(n)}) \|^2\right]$은 디바이스들이 동일한 경향의 데이터를 가지는 독립 항등 분포 (independent and identically distribution) 환경에서 $\mathbb{E}\left[\| \boldsymbol{g}(\boldsymbol{w}^{(n)}) \|^2\right]$으로 치환된다. 따라서 식 8과 9를 이용하여 식 7를 정리하면 아래와 같다.

$$\begin{aligned}&\mathbb{E}\left[\|\hat{\boldsymbol{g}}(\boldsymbol{w}^{(n)}) - \nabla F(\boldsymbol{w}^{(n)})\|^2\right] \\ &\le \beta' \| \nabla F(\boldsymbol{w}^{(n)}) \|^2 + \sigma'^2\end{aligned} \tag{10}$$

이때, β'과 σ'^2은 $\beta' = (1 + \mathbb{E}_j[c_j])\beta + \mathbb{E}_j[c_j]$와 $\sigma'^2 = (1 + \mathbb{E}_j[c_j])\sigma^2$를 만족하며, 양자화 과정으로 인하여 기존 SGD의 분산이 증폭되는 것을 확인할 수 있다. 또한 양자화 이질성을 반영하여 학습에 참여하는 디바이스의 양자화 비율에 영향을 받는 것 또한 식 10을 통해 확인할 수 있다. 따라서 이러한 양자화 이질성이 존재하는 환경에서 이를 고려한 로컬 학습을 새롭게 최적화할 필요가 있다.

로컬 학습을 최적화하는 기존 연구[3]의 Error upper-bound analysis를 통해, T_k번째 wall clock

time에서 global loss function은 아래와 같은 상한을 갖는다.

$$\frac{2[F(\boldsymbol{w}^{(n_k)})-F_{inf}]}{\eta T_k}\left(Y_k+\frac{D_k}{\tau_k}\right) + \frac{\eta L\sigma'^2}{|\mathbb{J}|}+\eta^2L^2\sigma'^2(\tau_k-1) \tag{11}$$

이때, T_k는 wall-clock time을 업데이트하는 k번째 시간 주기를, Y_k와 D_k는 각각 해당 주기의 연산 시간과 전송 시간을, τ_k는 해당 주기의 로컬 학습 횟수를 의미한다. 따라서 시간 주기 T_k에 해당되는 global round의 집합 $\mathbb{N}_k$는 $\{n_k, n_k+1, ..., n_{k+1}-1\}$로 구성되며, 이때 n_k는 k번째 wall-clock time에 시작 라운드를 의미한다.

식 11을 이용하여 k번째 wall-clock time T_k에서 로컬 학습 횟수를 최적화하면 아래와 같다.

$$\tau^{(n)}=\tau_k=\sqrt{\frac{2[F(\boldsymbol{w}^{(n_k)})-F_{inf}]D_k}{\eta^3L^2T_k\sigma'^2}} \tag{12}$$

해당 주기에서 로컬 학습 횟수는 $n\in\mathbb{N}_k$를 만족하는 모든 라운드 n에 대하여 τ_k로 동일하다. 이후 다음 시간 주기 T_{k+1}에서의 로컬 학습 횟수 τ_{k+1}은 점화식을 이용하여 간단하게 $\tau_{k+1}=\tau_k\sqrt{\frac{F(\boldsymbol{w}^{(n_{k+1})})-F_{inf}}{F(\boldsymbol{w}^{(n_k)})-F_{inf}}}$로 표현할 수 있다.

2.4 업로딩 지역 최적화

앞 절에서 제안한 업로딩 지역에서 양자화 이질성을 고려한 로컬 학습을 최적화하였다. J. Wang et al.의 연구[3]에서 동적 로컬 학습 횟수를 활용한 최종 학습 loss의 upper bound는 다음과 같이 정리된다.

$$\sum_{k=0}^{K}\phi(\tau_k) = \frac{2[F(\boldsymbol{w}^{(0)})-F_{inf}]}{\eta\sum_{k=0}^{K}\tau_k}+\frac{\eta L\sigma^2}{|\mathbb{J}|} +\eta^2L^2\sigma^2\left(\frac{\sum_{k=0}^{K}\tau_k^{\,2}}{\sum_{k=0}^{K}\tau_k}-1\right) \tag{13}$$

식 13은 로컬 학습에 관한 1, 3번째 항과 학습에 참여하는 디바이스 수에 관한 2번째 항으로 이루어지며, 로컬 학습 횟수를 의미하는 τ_k는 식 12를 통하여 기준 전송 시간을 의미하는 D_k에 비례한다. 이는 전송 시간이 오래 걸리는 경우 제한된 wall-clock time 안에 학습 로스를 빠르게 감소시키기 위해 상대적으로 시간이 적은 로컬 학습이 더 많이 필요하기 때문이다. 그리고 기준 전송 시간 D_k는 식 6에 의하여 업로딩 지역의 크기에 관한 값으로 정리된다. 또한 학습에 참여하는 디바이스의 수 역시 식 2에 의하여 업로딩 지역의 크기를 이용하여 정의된다.

따라서 학습 성능인 loss의 upper bound를 의미하는 식 13의 모든 항이 업로딩 지역의 크기에 밀접한 연관이 있는 것을 알 수 있으며, 이를 두 개의 서로 다른 업로딩 지역인 small zone과 large zone에서의 영향을 아래 표 1에 정리하였다.

업로딩 지역이 증가하게 되면 많은 디바이스가 학습에 참여하여 더 많은 데이터 셋을 학습에 활용할 수 있어 학습 성능이 향상된다. 하지만 이렇게 증가한 디바이스들이 학습에 참여하기 위한 기준 전송 시간 역시 증가하게 되고, 이는 제한된 wall clock time 안에 학습 loss를 효과적으로 감소시키기 위하여 더 많은 로컬 학습 횟수가 필요하다. 이러한 많은 로컬 학습 횟수는 기존 연구[3]를 통하여 학습 성능의 손해를 미치는 것을 알 수 있다. 이러한 trade-off를 통하여 학습 성능을 최대로 하도록 업로딩 지역의 크기를 본 절에서 최적화하고자 한다.

학습 성능의 의미하는 식 13을 업로딩 지역의 크기 r_{zone}과 로컬 학습 횟수 τ_k를 변수로 갖도록 표현하면 아래와 같다.

$$\sum_{k=0}^{K}\phi(\tau_k, r_{zone}) \tag{14}$$

로컬 학습 횟수를 최적화하기 위하여 업로딩 지역의 크기 r_{zone}을 활용하므로, 학습 로스를 최소로 하는 업로딩 지역의 크기를 구하는 최적화 문제는 다음과 같이 정의된다.

표 1. 업로딩 지역에 따른 학습 성능
Table 1. Learning Performance about the size of Uploading Zone

	N_{zone}	τ_k	1st	2nd	3rd
Small Zone	More	Big	↓	↓	↑
Large Zone	Less	Small	↑	↑	↓

$$\min_{r_{zone}}\left\{\sum_{k=0}^{K}\min_{\tau_k}\phi(\tau_k, r_{zone})\right\}$$
$$\text{s.t.} \quad r_{zone} \geq 0 \qquad (15)$$
$$D_k + Y_k \leq T_k$$
$$\tau_k = a_k \times {r_{zone}}^{\alpha/2}$$

a_k는 식 12의 D_k항에 기준 전송 시간을 의미하는 식 6을 대입하여 r_{zone}항을 제외한 나머지 값을 의미한다. 제약 조건으로 업로딩 지역에 관한 조건, 그리고 전송 시간과 연산 시간의 합은 wall-clock time보다 작아야 하며, 최적의 로컬 학습 횟수로 구성된다. 식 13을 제안하는 업로딩 지역 환경을 적용하면 식 15의 최적화 문제는 아래와 같이 표현된다.

$$\frac{2[F(\boldsymbol{w}^{(0)}) - F_{inf}]}{\eta {r_{zone}}^{\alpha/2}\sum_{k=0}^{K}a_k} + \frac{\eta L \sigma'^2}{\mu {r_{zone}}^2}$$
$$+\eta^2 L^2 \sigma'^2 {r_{zone}}^{\alpha/2}\left(\frac{\sum_{k=0}^{K}{a_k}^2}{\sum_{k=0}^{K}a_k} - 1\right) \qquad (16)$$
$$= C_1 (r_{zone})^{-\alpha/2} + C_2 (r_{zone})^{-2}$$
$$+ C_3 (r_{zone})^{\alpha/2} - \eta^2 L^2 \sigma'^2$$

이때, μ는 디바이스 분포 밀도를 의미하며 $\mu = N_{total}/{r_{end}}^2$으로 정의된다. 따라서 식 16이 r_{zone}에 대하여 convex 함을 보이기 위하여 2번 미분하면 path loss α가 $\alpha \geq 2$를 만족하는 경우 항상 convex 함을 간단하게 증명할 수 있다. 따라서 최적의 업로딩 지역의 크기 r^*_{zone}은 다음과 같이 non-closed form으로 표현된다.

$$\alpha (r^*)^2 (C_3 (r^*)^{\alpha/2} - C_1 (r^*)^{-\alpha/2}) = 4C_2 \qquad (17)$$

C_1, C_2, C_3 상수들은 학습 파라미터 L과 σ'^2 같은 간단하게 정의되지 않은 값들을 사용하여 표현하기 때문에 그 값을 정확하게 구하는 것이 어렵다는 한계가 존재한다. 따라서 이를 효과적으로 해결하기 위해 P. Liu et al.[5]에서 활용한 joint data-and-model-driven fitting 기법을 적용하여 우선 두 개의 업로딩 지역 r_{small}, r_{large}에서 학습을 진행하여 얻은 결과를 활용하면 역으로 추정하는 것이 가능하다. 이렇게 얻은 상수 C_1, C_2, C_3를 식 17에 대입하는 것을 통해 학습 파라미터의 한계를 해결하여 최적의 업로딩 지역 r^*_{zone}을 얻을 수 있다.

Ⅲ. 실 험

$N_{total} = 100$개의 디바이스가 uniform distribution으로 $r_{end} = 100m$에 디바이스 지역에 분포된 환경을 가정한다. 이때, 서로 다른 크기의 업로딩 지역 Small Zone과 Large Zone은 각각 30m, 60m의 크기를 갖는다. Small Zone에서의 기준 데이터 전송률은 2Mbps이며, 각 디바이스의 전송률은 위치에 따른 PL을 고려하여 계산되며, PL exponent α는 2로 가정하였다.

학습에는 MNIST와 Fashion-MNIST (FMNIST) 데이터 셋을 활용하였으며, 각 디바이스는 600개의 데이터를 학습데이터로 활용한다. 학습 모델로는 기존 연구[8]에서 사용한 [784, 400, 400, 10]의 FNN 네트워크로 구성된다. 학습은 미니 배치를 활용한 SGD 방법으로 진행되며, 미니 배치 ξ의 크기는 8로, 1,000개의 global round 동안 진행되며 learning rate η는 0.01이다. 최초의 로컬 학습 횟수 τ_0은 simple grid search를 활용하여 학습에 사용되며, 로컬 학습 횟수를 업데이트하는 wall-clock time 주기 T는 60초로 만약 다음 주기의 로컬 학습 횟수가 감소하지 않는 경우 수동적으로 이전 주기의 로컬 학습 횟수의 절반으로 설정된다.

실험 과정은 다음과 같다. 우선 MNIST와 FMNIST 데이터가 고르게 분포된 iid 환경에서 업로딩 지역 크기에 따른 loss 변화를 통해 최적의 업로딩 지역을 확인하고, II의 4절에서 제안하는 방법을 통해 얻은 optimal zone과 기존 small zone과 large zone에서 학습 성능을 각각 iid 환경과 non-iid 환경에서 비교하였다. 이때, non-i.i.d 환경은 연구[7], [8], [12]와 동일한 환경을 가정하여 진행하였다.

3.1 업로딩 지역의 크기에 따른 학습 성능

업로딩 지역에서의 학습 성능의 upper-bound analysis를 활용하면 식 16과 같이 표현된다. 이를 각각 MNIST와 FMNIST 데이터가 iid하게 분포된 환경에서 업로딩 지역의 크기에 따른 최종 Test loss 값을 나타내면 그림 3과 같다.

그림 3을 통하여 일반적으로 업로딩 지역의 크기가 증가함에 따라 더 많은 디바이스들이 학습에 참여하기 때문에, 학습 측면에서 데이터 셋의 증가로 인하여 loss 값이 작아져 학습 성능이 좋아지는 것을 알 수 있

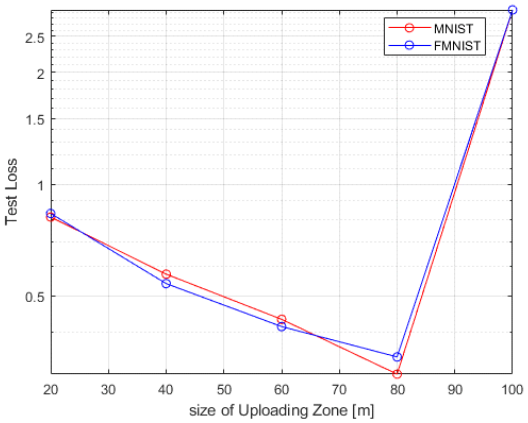


그림 3. 업로딩 지역 크기에 따른 Test loss 값 변화(iid MNIST, FMNIST 데이터 셋 환경)
Fig. 3. Test Loss about the size of Uploading Zone (iid MNIST and FMNIST data set)

다. 하지만 특정 임계값(80m, optimal zone) 이상의 uploading zone에서는 이러한 학습 데이터 증가로 인한 학습 성능 향상보다 제한된 wall-clock time 안에 loss 값을 줄이기 위해 사용되는 더 많은 로컬 학습 횟수로 인하여 학습 성능이 나빠지는 것을 확인할 수 있다. 이러한 경향성은 앞서 학습 성능을 최대로 하는 업로딩 지역 최적화 문제인 식 15를 통해 확인한 것과 일치하며, 식 17을 통해 구한 최적의 업로딩 지역의 크기는 MNIST 데이터와 FMNIST 데이터 환경에서 80m 지역이다.

3.2 MNIST 데이터 셋

제안하는 업로딩 지역에서 0부터 9까지의 손글씨 숫자 이미지 데이터인 MNIST 데이터를 활용하여 학습하였다. 디바이스들이 보유한 데이터의 경향이 iid한 환경과 non-iid한 환경에서 학습 성능을 비교해보면 그림 4, 5와 같다. 앞서 정의한 Small Zone과 Large Zone 그리고 최적의 업로딩 지역에서 전송 시간과 연산 시간을 합친 실제 학습 시간에 따른 학습 정확도를 분석하였다. 이때, 최적의 업로딩 지역의 크기 r^*_{zone}는 앞선 실험 결과에서 확인한 80m 지역이다.

그림 4와 5의 결과를 보면 제안하는 학습 성능을 최대로 하도록 최적화한 optimal zone에서 iid와 non-iid한 환경 모두에서 가장 학습 성능이 높게 나오는 것을 확인할 수 있다. 디바이스들이 iid한 데이터를 보유할 때와 비교하여 non-iid한 데이터를 가짐으로 인하여 test acc 정확도가 상대적으로 불안정하게 증가하는 것을 확인할 수 있지만, 성공적으로 학습이 수렴하는 것을 알 수 있다. 하지만 상대적으로 더 적은

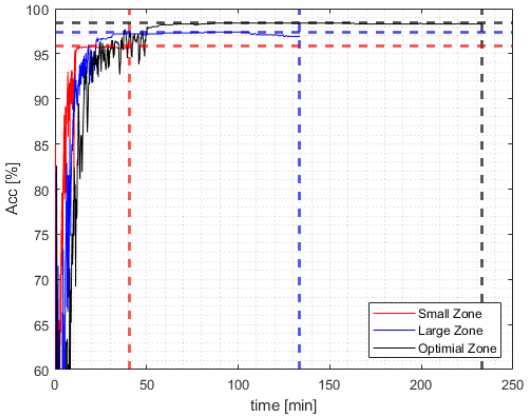


그림 4. 업로딩 지역에서의 학습 성능 비교(iid MNIST 데이터 셋 환경)
Fig. 4. Learning Performance according to the Uploading Zone (iid MNIST data set)

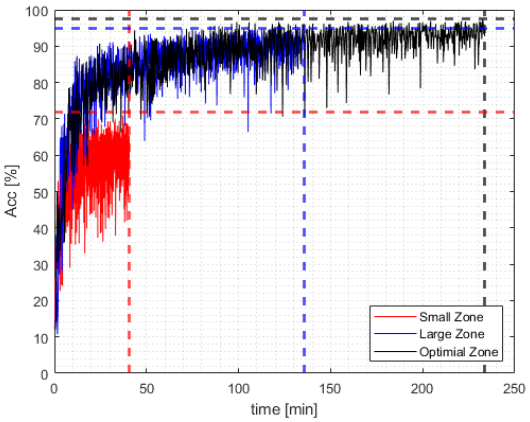


그림 5. 업로딩 지역에서의 학습 성능 비교(non-iid MNIST 데이터 셋 환경)
Fig. 5. Learning Performance according to the Uploading Zone (non-iid MNIST data set)

디바이스가 학습에 참여하는 small zone에서는 iid한 경우와 비교하여 수렴된 학습 정확도 측면에서 성능 하락이 존재한다.

3.3 Fashion-MNIST 데이터 셋

앞서 확인한 MNIST 데이터 대신 더 복잡한 데이터인 10개의 의류 이미지 데이터인 FMNIST 데이터를 활용하여 학습을 진행하였다. 이때, 디바이스들이 iid한 데이터를 보유한 환경과 non-iid한 데이터를 보유한 환경에서 학습 성능을 비교하면 각각 그림 6, 7과 같다. 최적의 업로딩 지역 r^*_{zone}은 80m 지역이며, 이를 small zone과 large zone에서의 학습 정확도를 비교하였다.

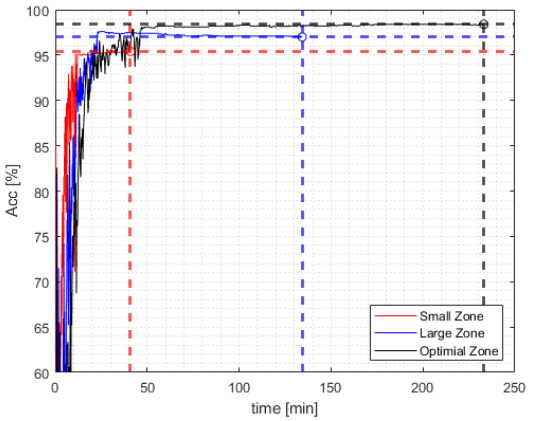


그림 6. 업로딩 지역에서의 학습 성능 비교(iid FMNIST 데이터 셋 환경)
Fig. 6. Learning Performance according to the Uploading Zone (iid FMNIST data set)

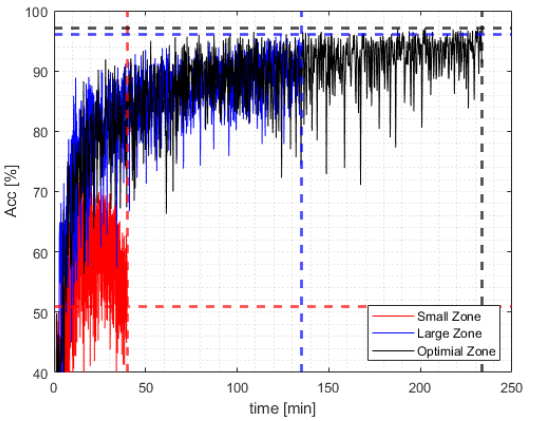


그림 7. 업로딩 지역에서의 학습 성능 비교(non-iid FMNIST 데이터 셋 환경)
Fig. 7. Learning Performance according to the Uploading Zone (non-iid FMNIST data set)

그림 6과 7의 결과를 통해 디바이스들의 데이터 경향성의 상관없이 항상 최적의 업로딩 지역에서 학습 정확도가 가장 높은 것을 확인할 수 있다. 또한 FMNIST 데이터의 복잡성으로 인하여 MNIST 데이터를 활용한 결과와 비교하여 상대적으로 낮은 정확도를 갖는 것 역시 확인할 수 있다. 디바이스들의 non-iid 데이터 경향은 적은 디바이스가 학습에 참여하는 small-zone에서 치명적이며, 너무 적은 디바이스로 인하여 학습 측면에서 사용되는 데이터들의 부족으로 해당 지역에서 학습이 충분히 수렴하지 않는 것을 알 수 있다.

Ⅳ. 결 론

본 논문에선 디바이스가 분포된 환경에서 업로딩 지역을 정의하고, 디바이스 위치를 고려한 양자화 기법을 적용하여 양자화 이질성이 있는 환경을 고려하였다. 이러한 환경에서 양자화 이질성으로 인한 오류 증폭에 따른 로컬 학습 횟수를 최적화하여 학습 성능을 분석하였다. 그 결과, 업로딩 지역이 증가하면 학습에 더 많은 디바이스가 참여하여 학습 정확도 측면에서 이득이 존재하지만, 기준 전송 시간 역시 증가하여 학습 수렴이 오래 걸리는 것을 확인하였다. 반면에 작은 업로딩 지역에서는 상대적으로 적은 디바이스가 학습에 참여하기 때문에 낮은 학습 정확도를 가지지만 기준 전송 시간이 짧으므로 학습이 빠르게 수렴하는 것을 알 수 있다. 그리고 학습 성능을 최대로 하도록 업로딩 지역의 크기를 최적화하였다. 일정 크기 이상의 업로딩 지역에서는 제한된 시간 안에 학습 loss를 줄이기 위해 사용하는 많은 로컬 학습으로 인하여 학습 성능이 감소하는 것을 확인할 수 있다.

References

[1] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, “Federated learning in mobile edge networks: A comprehensive survey,” *IEEE Commun. Surv. & Tuts.*, vol. 22, no. 3, pp. 2031-2063, 3rd quarter 2020. (https://doi.org/10.1109/COMST.2020.2986024)

[2] J. Wang and G. Joshi, “Cooperative SGD: A unified framework for the design and anaylsis of communication-efficient SGD algorithms,” *J. Mach. Learn. Res.*, vol. 22, no. 213, pp. 1-50, 2021.

[3] J. Wang and G. Joshi, “Adaptive communication strategies to achieve the best error-runtime trade-off in local update SGD,” in *Proc. Conf. Mach. Learn. and Syst.*, vol. 1, pp. 126-145, 2019.

[4] D. Alistarh, D. Grubic, J. Li, R. Tomika, and M. Vojnovic, “QSGD: Communication-efficient SGD via gradient quantization and encoding,” in *Proc. Advances in NIPS*, pp. 1709-1720, 2017.

[5] P. Liu, J. Jiang, G. Zhu, L. Cheng, W. Jiang,

W. Luo, Y. Du, and Z. Wang, "Training time minimization in quantized federated edge learning under bandwidth constraint," in *2022 IEEE WCNC*, pp. 530-535, Apr. 2022. (https://doi.org/10.1109/WCNC51071.2022.9771723)

[6] D. Jhunjhunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, "Adaptive quantization of model updates for communication - efficient federated learning," *2021 IEEE ICASSP*, pp. 3110-3114, 2021. (https://doi.org/10.1109/ICASSP39728.2021.9413697)

[7] C. Shen and S. Chen, "Federated learning with heterogeneous quantization," *2020 IEEE/ACM SEC*, Nov. 2020. (https://10.1109/SEC50012.2020.00060)

[8] S. Chen, C. Shen, L. Zhang, and Y. Tang, "Dynamic aggregation for heterogeneous quantization in federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6804-6819, Oct. 2021. (https://doi.org/10.1109/TWC.2021.3076613)

[9] M. K. Nori, S. Yun, and L-M. Kim, "Fast federated learning by balancing communication trade-offs," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5168-5182, Aug. 2021. (https://doi.org/10.1109/TCOMM.2021.3083316)

[10] S. H. Jeon and D. I. Kim, "A study on the federated learning performance analysis about uploading zone," in *Proc. Symp. KICS*, pp. 1137-1138, Jeju Island, Korea, Jun. 2022.

[11] S. H. Jeon and D. I. Kim, "A study on the local update optimization on quantization heterogeneity," in *Proc. Symp. KICS*, pp. 810-811, Pyeongchang, Gangwon Province, Korea, Feb. 2022.

[12] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, pp. 1273-1282, Apr. 2017.

전 성 훈 (Seong Hoon Jeon)

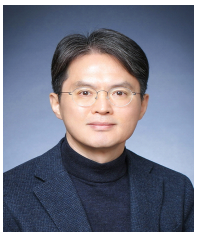
2021년 2월 : 성균관대학교 전자전기공학과 졸업

2021년 3월~현재 : 성균관대학교 전자전기컴퓨터공학과 석사 과정

<관심분야> IoT, 연합 학습, 양자화

[ORCID:0000-0002-6267-5981]

김 동 인 (Dong In Kim)



1980년 2월 : 서울대학교 전자공학과 졸업 (학사)

1987년 12월 : Electrical Engineering, University of Southern California 졸업 (석사)

1990년 12월 : Electrical Engineering, University of Southern California 졸업 (박사)

2007년 9월~현재 : 성균관대학교 전자전기공학과 교수

<관심분야> 무선통신, IoT, 무선전력전송, 연합학습

[ORCID:0000-0001-7711-8072]