Journal of KIIT. Vol. 23, No. 9, pp. 1-8, Sep. 30, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.9.1

연합학습에서 오려낸 양자화 적용에 대한 성능 분석

임 승 호*

Performance Analysis of Clipped Quantization in Federated Learning

Seung-Ho Lim*

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2021R1F1A1048026). 이 연구는 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것임

요 약

연합학습은 학습을 개별 클라이언트에 분산시켜 학습연산의 오버헤드는 분산되지만, 중앙 서버에 모델 파라미터를 전송하는 오버헤드가 증가한다. 본 연구에서는 연합학습에서 클라이언트와 중앙서버간 전송되는 파라미터의 양을 줄이기 위해서 오려낸(Clipped) 양자화를 파라미터에 적용하여 본다. 오려낸 양자화(Clipped Quantization)는 연합학습 시스템에서 클라이언트가 매 라운드 시 업데이트된 파라미터를 양자화하여 중앙서버에 보낼 때, 양자화 값의 분포를 파악하여 양자화된 파라미터의 양 끝부분을 일정 비율로 클리핑한 후 이 값을 압축하여 보내는 방식이다. 이러한 오리기 방법은 양자화된 값 중에서 유효한 값을 가지는 범위가 줄어듦으로써 압축률을 높일 수 있으므로 전송량을 줄일 수 있다. 실험결과, 딥러닝 모델에서 2~5%의 정확도 감소 대비 30~40% 에폭 당 데이터 전송량이 줄어듦을 확인할 수 있었다.

Abstract

Federated learning distributes learning to individual clients, so the overhead of learning operations is distributed, but the overhead of transmitting model parameters to the central server increases. In this paper, clipped quantization is applied to parameters to reduce the amount of parameters transmitted between the client and the central server in federated learning. Clipped Quantization is a method in which, when a client quantizes the updated parameters for each round in a federated learning system and sends them to the central server, the distribution of the quantized values is identified, and the ends of the quantized parameters are clipped by a certain ratio and then these values are compressed and sent. This clipping method can reduce the amount of data transmitted because the compression ratio can be increased by reducing the range of valid values among the quantized values. Experimental results have confirmed that the amount of data transmitted is reduced relative to the decrease in precision in various deep learning models.

Keywords

federated learning, quantization, clipping, parameter transmission

* 한국외국어대학교 컴퓨터공학부 교수
- ORCID: http://orcid.org/0000-0003-3096-0785

· Received: Mar. 20, 2025, Revised: Jul. 28, 2025, Accepted: Jul. 31, 2025
· Corresponding Author: Seung-Ho Lim
Division of Computer Engineering, Hankuk University of Foreign Studies, Korea
Tel.: +82-31-330-4704, Email: lim.seungho@gmail.com

Ⅰ. 서 론

최근, 딥러닝 기술을 포함한 인공지능 기술의 발전과 더불어 인공지능 학습과 추론을 위해서 대용량의 데이터에 대한 수집, 저장 및 가공하여 이를 인공지능 알고리즘으로 실행하는 응용 애플리케이션 및 시스템의 사용성이 점점 증대하고 있으며[1] 의료분야나 개인 일상생활 등을 포함한 다양한 개인정보 분야로 확장되면서 사적인 데이터에 의한 개인정보 유출에 대한 위험성이 증대되었다[2].

인공지능 학습과 추론을 수행하기 위해서 개인정보가 포함된 데이터를 인공지능 서버로 보내는 과정에서 개인정보의 유출을 방지하기 위해서 연합학습 방법이 제안되었다[2][3]. 연합학습은 개인정보가 포함된 데이터를 중앙서버에서 수집하여 인공지능 연산을 통한 학습을 수행하는 대신 개별 클라이언트에서 인공지능 학습을 수행한 후 학습 결과인 모델의 파라미터를 중앙서버에서 수집하여 모델의 평균화 작업을 통해서 업데이트하고, 업데이트된 모델을 다시 개별 클라이언트에게 되돌려준다. 개별 클라이언트는 서버로부터 전송받은 모델 파라미터를 이용하여 다시 학습을 수행한다. 이러한 반복과정을 통해서 학습을 수행함으로써 개인정보를 포함한 데이터의 공유 없이 학습률을 높이는 방법이며, 이렇게 클라이언트와 중앙서버가 학습과 모델 업데이트를 반복해서 수행하는 분산 학습 시스템이라고 할 수 있다.

연합학습은 데이터의 공유 없이 인공지능 모델의 파라미터만 공유함으로써 사적인 데이터의 유출에 대한 문제점을 해결할 수 있다. 그러나, 매 학습 라운드 시에 클라이언트들은 파라미터를 서버와 송수신함으로써 발생하는 네트워크 전송의 오버헤드가 높으며, 이러한 파라미터 전송 오버헤드를 낮추는 효율적인 방법이 필요하다. 또한 파라미터만으로도 학습에 사용한 데이터를 유추하는 파라미터 공격 알고리즘에 대한 대처방안도 요구된다.

연합학습에서 클라이언트의 파라미터 전송량을 줄이는 방법으로 파라미터 양자화를 적용할 수 있으며, 기존의 많은 연합학습 연구에서 이를 적용하였다[4]-[7]. 일반적으로 인공지능 모델의 파라미터는 32 비트 부동소수점인데, 이러한 부동소수점을 비트 수가 작은 정수형으로 변환하여 파라미터의 크기 및 인공지능 연산의 복잡도를 낮추는 방식이다. 일반적으로 32비트 부동소수점을 8비트나 그 이하의 비트로 양자화하는 방식이 많이 적용된다. 그러나 연합학습에서는 다수의 클라이언트에서 매 학습 라운드 시 파라미터를 송수신하기 때문에 네트워크 전송량을 더욱더 줄이는 방식이 필요하다.

본 논문에서는 연합학습에서 클라이언트와 중앙서버 간 전송되는 파라미터의 양을 줄이기 위해서 오려낸(Clipped) 양자화를 파라미터에 적용하여 그 성능을 분석하여 보았다. 오려낸 양자화(Clipped quantization)는 연합학습 시스템에서 클라이언트가 매 라운드 시 업데이트된 파라미터를 양자화하여 중앙서버에 보낼 때, 양자화 값의 분포를 파악하여 양자화된 파라미터의 양 끝부분을 일정 비율로 오려낸(Clipping) 후 이 값을 압축하여 보내는 방식이다. 이렇게 양자화된 파라미터의 일정 부분을 오려냄으로써 송수신 양을 줄일 수 있다. 오려진 비율이 높을수록 파라미터의 유효한 비트 범위가 줄어들 수 있으므로 압축률이 높아질 수 있어 파라미터 전송량을 줄일 수 있지만, 학습된 파라미터값을 잘라내는 것으로 인해서 학습률이 낮아질 수 있다. 우리는 본 논문에서 연합학습 시스템에서 오려낸 양자화를 구현하여 보고, 이를 다양한 인공지능 딥러닝 모델에 적용하여 보고 실험을 수행하여 그 성능을 분석하여 보고, 오려낸 양자화를 적용한 연합학습 인공지능 모델에서 적은 데이터 전송량 대비 목표의 학습률을 달성할 수 있는 연합학습 시스템의 효율성을 높일 수 있을 것으로 본다.

Ⅱ. 배경 및 관련 연구

연합학습은 분산 시스템에서 딥러닝 모델과 같은 인공지능 모델을 학습시키는 분산 학습 방법으로써 모델의 학습을 수행하는 다수의 클라이언트와 개별 학습 된 클라이언트들의 모델을 통합하여 업데이트하는 서버로 구성된다. 연합학습 시스템에서 학습과 모델의 업데이트는 라운드마다 클라이언트와 서버의 파라미터 송수신을 동반한다. 라운드마다 클라이언트는 자체 수집한 데이터에 대해서 모델 학습을 수행하고, 수행된 결과 파라미터를 서버에 송신한다. 서버는 모든 클라이언트 혹은 각 라운드에서 지

정된 특정 클라이언트로부터 파라미터를 수신받아서 파라미터 업데이트를 수행한다. 파라미터를 업데이트하는 알고리즘은 FedAVG[8][9]이 원천적인 알고리즘인데, 이는 수신받은 파라미터들의 평균으로 새로운 파라미터를 구한 후, 이 값을 다시 클라이언트들에게 전송한다. 이후 FedAVG에 기반한 여러 가지의 확장적인 알고리즘 방식들이 연구되었으며 FedAPA[10], FedDUAP[11], pFedGPA[12] 등이 있다. 본 논문은 원천적이고 기본적인 FedAVG 알고리즘을 파라미터 업데이트 알고리즘으로 적용한다.

매 학습 라운드 시 서버와 클라이언트 사이의 파라미터 송수신 오버헤드를 줄이기 위해서 다양한 방식이 연구되었으며[4]-[7][13] 주요한 기술이 양자화를 적용하는 방법이다. 양자화는 일반적으로 부동소수점 표현인 Floating Point 32비트로 구성된 신경망 모델의 파라미터를 정수형 데이터인 Integer로 압축하는 방식이 대표적이다[4][6][13]. 또한, 연합학습에서는 라운드마다 파라미터의 전송을 동반하며 학습 시 업데이트되는 파라미터의 변화량이 많지 않음을 고려하여 파라미터의 양자화 압축률을 높이기 위해서 파라미터 그래디언트(Gradient) 양자화를 수행한다[5][7]. Gradient 양자화는 학습에 따른 파라미터의 변화량에 대한 양자화를 수행하기 때문에 변화량에 대한 표현 가능한 Integer 비트 수를 많이 줄일 수 있다[4]. 그러나, 여전히 연합학습에서는 라운드마다 송수신하는 파라미터의 전송량을 줄이는 방안이 요구된다. 또한, 일정한 크기와 패턴의 파라미터를 송수신하는 것은 파라미터의 패턴을 유출함으로써 Inversion Attack[14]-[16]과 같은 데이터 유출과 보안에 문제점이 있을 수 있다.

양자화 파라미터의 압축률을 높이는 방법으로 가지치기(Pruning)[9]나 오리기(Clip)[4][6]과 같이 네트워크 연산 및 학습에 영향을 많이 미치지 않는 파라미터를 제거하거나 0으로 설정함으로써 파라미터의 희소 행렬(Sparse matrix) 화를 높이는 방법을 적용할 수 있다. 이러한 방법은 뉴럴 네트워크의 학습률에 영향을 많이 미치지 않는 파라미터값들을 제거함으로써 압축률을 높이지만, 너무 많은 파라미터 값을 제거할 때 학습률이 떨어지기 때문에[4][6][16], 연합학습에서 네트워크 전송량과 학습률 향상과의 상관관계를 고려하여 적용할 필요가 있다. 본 논문에서는 오리기 기반 양자화를 적용한 연합학습 시스템을 구현하고 다양한 신경망에 대해서 이를 적용한 연합학습을 수행하여 보았다. 이를 통해서 학습률과 데이터 전송량의 상관관계를 분석한다.

III. 연합학습 시스템

3.1 연합학습 시스템 구성

그림 1은 Clipped 양자화를 구현 적용한 연합학습 시스템의 구성도를 나타낸 것이다. 그림에서와 같이 연합학습 시스템은 다수의 클라이언트와 중앙 서버로 구성되며 연합학습 시스템에서 신경망 모델은 공통으로 적용된다.

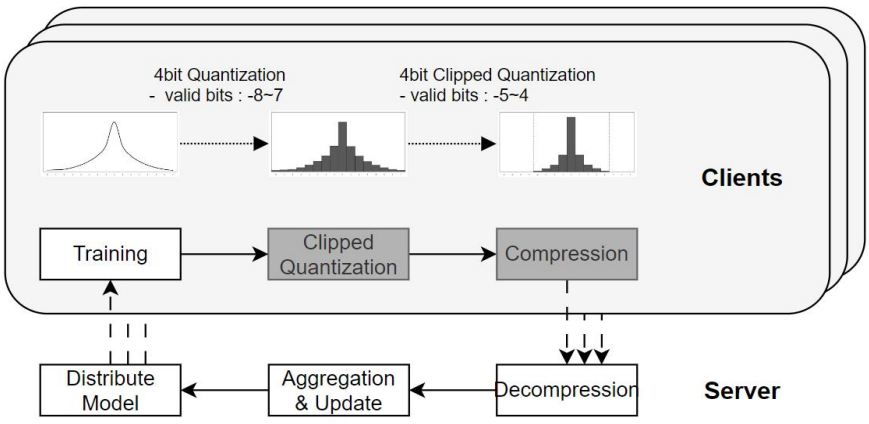

그림 1. 연합학습 시스템 구성도
Fig. 1. Federated learning system overview

라운드마다 자료수집과 학습은 클라이언트에서 수행하며, 학습된 모델의 파라미터는 서버에 전송된다. 서버는 클라이언트로부터 수신한 모델의 파라미터를 종합하여 업데이트한 후 이를 다시 각 클라이언트에게 되돌려준다.

각 클라이언트는 라운드마다 모델을 학습하고, 학습된 파라미터에 대한 Clipped 양자화 및 희소(Sparse) 파라미터의 압축을 수행한 후, 해당 파라미터를 서버에 전송하는 과정을 수행한다. 서버는 각 클라이언트로 전송받은 압축 파라미터를 푼 후에 평균을 내어 업데이트된 파라미터를 다시 클라이언트에게 전송한다. 이러한 과정을 원하는 학습률에 도달할 때까지 반복한다.

일반적으로 신경망 모델에서 파라미터값의 분포를 그래프로 도식화하면 가우시안 분포를 나타낸다[6]. 즉, 중앙값을 기준으로 파라미터의 값들이 가운데에 많이 분포하게 되고 양 끝으로 갈수록 값의 분포 비율이 낮아지며, 이는 양자화를 적용하더라도 당연히 같은 비율을 유지하게 된다. 이러한 분포도에서 값의 분포 비율이 낮은 양 끝부분을 일정 비율로 잘라내게 되면 이 부분이 유효값의 분포에서 사라지게 된다. 이렇게 유효값을 없앰으로써 파라미터 분포에서 유효한 값의 범위를 줄임으로써 압축률을 높일 수 있다. 그러나 또한 사라지는 부분으로 인해서 학습률에는 안 좋은 영향을 미칠 수 있다. 예를 들어서 4비트 양자화를 수행할 때 유효한 값을 가지는 비트 레벨의 수는 16레벨이 가능한데, 만약 양 끝 두 비트의 값을 제거하면 12레벨로 유효한 값을 표현할 수 있다. 이 경우 총표현 비트의 범위를 75%로 줄일 수 있게 된다.

3.2 Clipped 양자화 및 파라미터 압축

본 논문에서 적용하고 분석한 Clipped 양자화 방법은 다음과 같다. 먼저 각 클라이언트는 양자화되지 않은 32비트 부동소수점 파라미터를 이용하여 신경망 모델에 대한 학습을 수행하고 난 결과로 그래디언트를 생성한다. 생성된 그래디언트에 대해서 Clipped 양자화를 수행하기 위해서, 설정된 Clip Ratio를 사용한다. Clip Ratio는 전체 파라미터 분포에서 얼마의 비율로 양 끝 파라미터를 잘라낼 것인가에 대한 값이다. 일반적으로 양자화는 파라미터의 최소값과 최대값을 구하여 양자화 대상의 전체 구간을 설정하고, 각 파라미터값에 대해서 전체 구간의 위칫값으로 치환해 줌으로써 양자화를 수행한다. 또한 신경망에서는 각 레이어 별로 파라미터의 분포와 최소값 및 최대값의 분포가 다르므로 각 레이어별로 양자화를 수행한다.

Clipped 양자화는 그래디언트 양자화를 수행할 때 Clip Ratio 이용하여 파라미터의 최소값/최대값 폭을 Ratio만큼 수정하여 줌으로써 개별 파라미터값의 양자화 비율을 수정하는 방식으로 적용한다. 그림 2는 신경망의 계층별로 Clipped 양자화를 적용하는 방식을 도식화한 것이다. 즉, 신경망에서 계층별로 최소값과 최대값을 구한 후, 최대값과 최소값을 뺀 파라미터 구간에 Clip Ratio를 곱하여 주어 구간의 폭을 비율만큼 넓힌다. 그 후에 각 파라미터에 대해서 해당 구간에서 위치하는 위치값으로 양자화를 수행하여 준다. 이렇게 하여 주면 Clip Ratio만큼 양 끝에 위치하는 양자화 파라미터값이 사라지게 되며, 그 결과 유효한 값을 가지는 비트 레벨의 수를 줄여줄 수 있게 된다.

Clipped 양자화를 통해서 유효한 파라미터의 표현 비트 수를 줄임으로써 파라미터에서 유효한 데이터가 줄어들며 많은 영역이 0으로 표현되는 이른바 Sparse 파라미터의 형태로 표현되며 구조적으로 희소 행렬(Sparse matrix)로 볼 수 있다. 이러한 희소 행렬 형태의 모델 파라미터에 대해서 클라이언트와 서버 간의 전송량을 줄이기 위해서 압축을 수행해 주어야 한다. 희소 행렬의 압축을 위해서 COO(Coordinate List)로 표현하거나 CSR(Compressed Sparse Row) 방식[9]을 적용할 수 있다. 클라이언트는 모델의 각 계층에 대해서 희소 행렬에 대한 압축을 적용하여 파라미터의 용량을 줄여 압축된 파라미터를 서버에서 송신한다. 서버는 각 클라이언트로부터 수신받은 파라미터들에 대해서 압축해제를 통해서 원복시킨 후, 반 양자화(Dequatization)를 수행한다. 그 후에 파라미터를 평균하여 업데이트된 파라미터값을 생산한다. 이후 다시 이를 양자화하여 각 클라이언트에게 되돌려준다. 클라이언트는 수신받은 파라미터를 이용하여 기존에 가지고 있던 파라미터를 업데이트하고, 학습을 수행한다.

Ⅳ. 실험 및 실험 결과

우리는 Clipped 양자화 기반 연합학습 시스템을 오픈소스 연합학습 시스템[4]에 적용하여 다양한 실험을 진행하였다. 오픈소스 연합학습 시스템은 다수의 클라이언트와 서버를 모델링하여 구축한다. 클라이언트는 설정된 신경망 모델로 학습을 수행한 후 그래디언트 양자화 및 Clipping 과정을 거친 파라미터를 서버에 전송하고, 서버는 클라이언트로부터 송신 받은 파라미터를 FedAVG 알고리즘으로 종합하고 업데이트하여 다시 클라이언트로 되돌려주는 과정을 수행한다. 우리는 10개의 클라이언트를 모델링한 연합학습 시스템에 대해서 Alexnet 및 Resnet18 CNN 모델과 LSTM 모델을 적용하여 Clipped 양자화에 대한 실험을 수행하고 성능평가를 하였다. 실험은 각 신경망 모델에 대해서 양자화 비트 수를 8, 6, 4, 3으로 변화시켜 가면서 지정된 양자화 비트 수를 설정한 상황에서 Clipping Ratio를 No Clip, 0.9, 0.5, 0.1 Ratio로 설정을 변화해 가면서 실험을 수행하였다. Clipping Ratio는 No Clip은 Clipping을 적용하지 않은 것이며, 숫자가 낮아질수록 Clipping 적용을 위한 Min-Max 범위가 숫자의 비율의 역수배로 증가함을 의미한다. 즉 숫자가 작을수록 Clipping 양이 많아진다. 이렇게 각각 설정된 양자화 레벨과 Clipping Ratio에 대해서 100 epoch 동안 학습을 수행해 가면서 실험을 수행하였다. 실험 내용은 아래 표로 정리하였다.

표 1. 실험 데이터 세트
Table 1. Experimental data sets

Item	Contents	
# of Clients	10	
Epoch	100	
Model & Dataset	Alexnet	CIFAR-10
	Resnet18	CIFAR-10
	LSTM	shakespeare
Quant. Level	8bit, 6bit, 4bit, 3bit	
Clipping Ratio	No Clip, 0.9(Low Clipping) 0.5(Mid Clipping), 0.1(High Clipping)	
Federated	FedAVG	

그림 2는 Alexnet 모델에 대해서 양자화 레벨을 변화시키면서 각 양자화 레벨에서 Clip Ratio의 변화에 따른 학습률 정확도 실험 결과를 보여준 것이다. Alexnet, Resnet18, LSTM 모두 실험을 수행하였으나 경향성이 유사하여 Alexnet에 대해서 대표적으로 실험 결과를 나타내었다. 좌측 위에 도식화된 그림은 8bit 양자화 모델에 대해서 Clipping Ratio의 변화에 따른 학습률 정확도를 나타낸 것인데, 실험 결과에서 알 수 있듯이 8bit 양자화의 경우 Clipping Ratio에 큰 상관없이 학습률이 높아짐을 알 수 있다. 그러나 양자화 레벨이 낮아질수록, 즉 양자화 비트 수가 줄어들수록 Clipping Ratio에 따른 학습률 향상은 점점 좋지 않음을 알 수 있다. 그림의 우측 위에 나타난 양자화 레벨 6bit부터 Clipping Ratio에 따른 학습률에 차이가 나타난다. 즉, Clipping 양이 많아질수록 학습률은 떨어지며, 이는, 양자화 레벨이 4비트에서 3비트로 줄어들수록 그 차이는 점점 증가한다. 이는 Clipping에 의해서 오려지는 유효한 데이터가 늘어나므로 딥러닝 학습 효율성이 감소함에서 나타나는 것이다. 그러나 Clipping에 의해서 유효 비트 파라미터의 범위가 줄어들기 때문에 유효 파라미터의 감소에 따른 전송량의 감소를 가져올 수 있다.

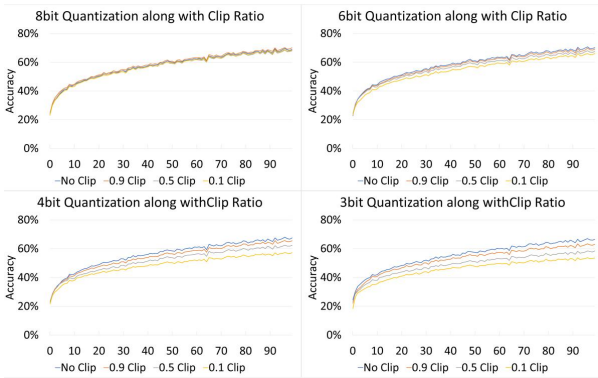

그림 2. Alexnet에서 양자화 레벨 8bits, 6bits, 4bits, 3bits에 대해서 Clip Ratio의 변화에 따른 학습률 정확도 실험 결과
Fig. 2. Results of learning rate accuracy experiments according to Clip Ratio changes for quantization levels of 8 bits, 6 bits, 4 bits, and 3 bits for the Alexnet

Clipping Ratio의 변화에 따른 딥러닝 모델의 유효 파라미터의 분포도를 파악하기 위해서 LSTM 8비트 양자화 모델과 Resnet18 4비트 양자화 모델에

대해서 Clipping Ratio의 변화에 따른 파라미터의 값의 분포도를 측정하여 그림 3에 나타내었다. 그림에서 x-축은 파라미터의 값을 나타내는 것이고, y-축은 각 파라미터의 개수 혹은 비율을 나타낸 것이다. 그림에서와 같이 Clipping Ratio의 숫자가 낮아질수록, 즉 Clipping 비율이 높아질수록 유효한 값을 가지는 파라미터의 범위가 줄어들면서 가운데로 뾰족하게 모이는 것을 확인할 수 있다. 이렇게 유효한 파라미터의 범위가 줄어들고, 특정 비트값이 높은 비율을 가지면 유효한 범위에 대해서만 특정 인코딩을 적용하거나 압축을 적용할 수 있으므로 높은 압축률을 달성하여 전송량을 줄일 수 있다. 특히 그림 3의 오른쪽 그림에 나타난 바와 같이 Resnet18 4비트 양자화 모델에 0.1 Clipping Ratio를 적용할 때 유효한 파라미터의 범위를 (-1, 0, 1)로 줄일 수 있으며 이는 Trinary 양자화의 레벨과 유사한 효과를 나타냄을 확인할 수 있다. 이 경우 앞선 실험 결과에서 유추할 수 있듯이 학습률을 조금 희생하더라도 더 높은 압축률을 달성할 수 있으므로 전송량 대비 학습률에서 효율성을 달성할 수 있다.

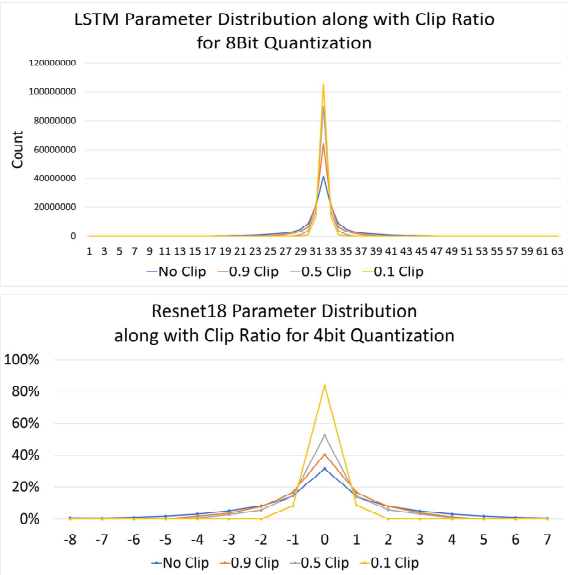

그림 3. LSTM-8bit 및 Resnet18-4bit 양자화 모델에서 Clipping Ratio의 변화에 따른 파라미터 유효 비트의 분포도

Fig. 3. Valid bits distribution according to the change of Clipping Ratio in LSTM-8bit and Resnet18-4bit quantization model

그림 4는 Alexnet 모델의 Cifar10 데이터 셋에 대해서 양자화 비트를 8비트에서부터 3비트까지 변화시켜 가면서 각 양자화 비트 레벨별로 Clipping Ratio를 변화해 가면서 100 에폭의 학습을 수행하여 측정한 정확도를 도식화한 결과이다. Alexnet 모델의 Cifar10 데이터셋의 경우에는 6비트 양자화 레벨에서부터 클리핑 비율에 따른 정확도 감소가 나타남을 확인할 수 있다. 클리핑 비율이 높을수록 파라미터의 압축률은 높아지지만, 정확도가 감소하므로 둘 사이의 상관관계를 적절히 조절할 필요가 있음을 알 수 있다. 같은 양자화 레벨의 경우 클리핑 비율에 따른 정확도의 감소가 무시할 만한 수준이라면 클리핑 비율을 높여서 파라미터 압축률을 높이면, 유사한 정확도와 학습률 대비 연합학습 시스템의 데이터 전송량 감소를 달성할 수 있다.

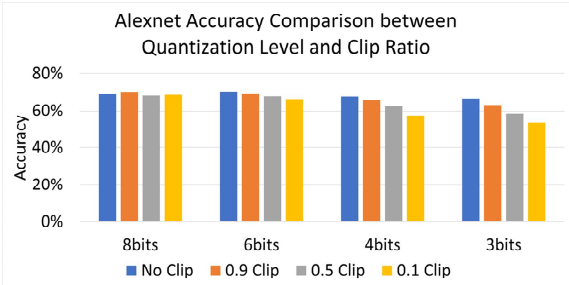

그림 4. Alexnet의 Cifar10 데이터셋에 대한 양자화 비트별 Clipping Ratio 변화에 따른 정확도 실험 결과

Fig. 4. Experiment results for accuracy according to changes in clipping ratio per quantization bit on the Cifar10 dataset on the Alexnet

그림 5는 Resnet18에 대해서 Cifar10 데이터셋에 대해서 양자화 레벨별로 Clipping Ratio를 변화시켜 가면서 실험한 결과에 대해서 정확도(왼쪽)와 데이터 전송량(오른쪽)을 각각 나타낸 것이다. 왼쪽 그래프에서 확인할 수 있는 것은 양자화 비트 수가 높을수록 Clipping Ratio에 대한 영향이 덜하며 양자화 비트 수가 낮아질수록 Clipping Ratio의 영향을 많이 받게 된다. 반면 오른쪽 그래프에서 확인할 수 있듯이, 같은 양자화 레벨에서 Clipping Ratio의 값에 따른 데이터 전송량에서 차이가 크게 남을 확인할 수 있다. 즉, 양자화 레벨이 같은 설정들에 대해서 클리핑 비율의 변화에 대해서 유사한 정확도를 보이는 설정들에 대해서 클리핑 비율을 높일수록 정확도의 유실 대비 데이터 전송량을 줄일 수 있음을 확인할 수 있다. 이러한 방식으로 양자화 레벨과

클리핑 비율들을 비교하여, 전송량과 학습률 및 정확도에 대한 최적의 효율적인 양자화 레벨 및 클리핑 비율을 결정한다. 예를 들어 양자화 레벨이 8bit 및 6bit의 경우 Clipping Ratio를 0.5로 설정할 때 정확도의 유실이 거의 없이 데이터 전송량을 에폭 당 34~38% 감소시킬 수 있음을 확인할 수 있다. Clipping Ratio를 0.1로 설정할 때 에폭당 전송량은 47~50% 정도 줄일 수 있으나, 정확도가 2~5% 감소하므로, 데이터 전송량과 정확도의 감소 폭을 고려할 필요가 있게 된다.

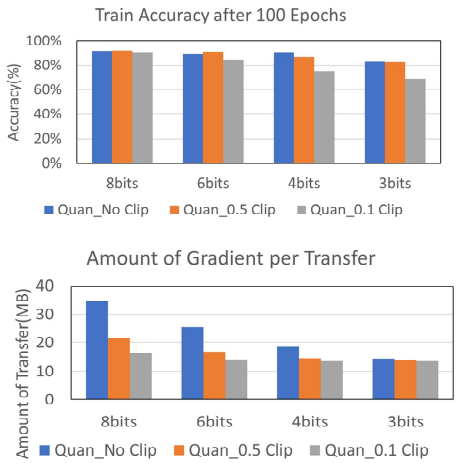

그림 5. Resnet18 모델의 Cifar10 데이터셋에 대한 양자화 비트별 Clipping Ratio 변화에 따른 정확도 및 전송량 비교 결과

Fig. 5. Experimental results on the relationship between accuracy and data transfer according to changes in the clipping ratio for the Cifar10 dataset of the Resnet18.

V. 결론 및 향후 과제

연합학습 시스템은 클라이언트의 로컬 데이터를 활용하여 학습한 파라미터를 서버에 공유하고, 이를 종합해 전역 모델을 갱신하는 보안 중심의 분산 학습 방식이다. 그러나 딥러닝 모델의 고도화로 파라미터가 증가함에 따라 학습 라운드마다 발생하는 서버-클라이언트 간 통신 오버헤드가 점차 심화하고 있다. 본 논문에서는 연합학습 시스템의 파라미터 전송량을 줄이고 효율적인 학습을 위한 클리핑 기반 파라미터 양자화 모델에 대한 구현과 실험을 수행하였다. 양자화된 파라미터값의 압축을 수행할 때 압축률을 높이기 위해서 양자화를 수행할 때 기본 범위를 늘려줌으로써 클리핑 효과 주도록 하면 양자화된 값 중에서 유효한 값을 가지는 범위가 줄어듦으로써 압축률을 높일 수 있다. 우리는 다양한 양자화 레벨에 대해서 클리핑 비율을 변화해 가면서 딥러닝 모델의 학습을 수행하여 그 실험 결과를 분석하였으며, 그 결과 각 양자화 레벨에서 다양한 클리핑 비율을 적용할 때 딥러닝 모델의 학습률과 데이터 전송량과의 관계가 있었고, 양자화레벨과 클리핑 비율에 따라서 정확도의 5% 감소를 희생하여 에폭 당 전송량이 30%이상 감소함을 확인하였다.

이를 통해서 딥러닝 모델에 따라서 최적화된 양자화 레벨과 클리핑 비율을 결정하여 효율적인 데이터 전송을 수행하는 연합학습 시스템을 구축할 수 있을 것으로 기대한다. 이를 기반으로 하여 향후 연구에서는 연합학습 시스템에 특정 딥러닝 모델을 적용할 때 해당 모델과 데이터셋에 대해서 양자화와 클리핑 비율에 따른 정확도 및 데이터 전송량과와 상관관계를 분석하여 실시간 동적 클리핑을 적용하는 연구를 수행할 계획이다.

References

[1] S. Yoo, K.-H. Lee, J. Park, S. J. Yoon, C. Cho, Y. J. Jung, and I. Y. Cho, "Trends in Deep Learning Inference Engines for Embedded Systems", Electronics and Telecommunications Trends, ETRI, Vol. 34, No. 4, pp 23-31, Aug. 2019. https://doi.org/10.22648/ETRI.2019.J.340403.

[2] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated Optimization: Distributed Machine Learning for On-Device Intelligence", arXiv preprint, arXiv:1610.02527, Oct. 2016. https://doi.org/10.48550/arXiv.1610.02527.

[3] A. Hard, et al., "Federated Learning for Mobile Keyboard Prediction", arXiv preprint, arXiv:1811.03604, Nov. 2018. https://doi.org/10.48550/arXiv.1811.03604.

[4] C. Zhang, S. Li, J. Xia, and W. Wang, "BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning", 2020 USENIX Annual Technical Conference (USENIX ATC 20), Boston, MA, USA, pp. 493-506, Jul. 2020.

[5] D′ EFOSSEZ, Alexandre; ADI, Yossi; SYNNAEVE, Gabriel, "Differentiable model compression via pseudo quantization noise", arXiv preprint arXiv:2104.09987, Apr. 2021. https://doi.org/10.48550/arXiv.2104.09987.

[6] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li., "Terngrad: Ternary gradients to reduce communication in distributed deep learning", Advances in neural information processing systems 30, Long Beach, California, USA, Dec. 2017.

[7] A. Reisizadeh, et al., "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization", International conference on artificial intelligence and statistics, PMLR, Online, Aug. 2020.

[8] H. B. McMahan, et al., "Communication-efficient learning of deep networks from decentralized data", Artificial intelligence and statistics, PMLR, Fort Lauderdale, FL, USA, Apr. 2017. https://doi.org/10.48550/arXiv.1602.05629.

[9] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data", Proc. 8th Int. Conf. Learning Representations (ICLR), Addis Ababa, Ethiopia, Apr. 2020. https://doi.org/10.48550/arXiv.1907.02189.

[10] Y. Sun, et al., "FedAPA: Server-side Gradient-Based Adaptive Personalized Aggregation for Federated Learning on Heterogeneous Data", arXiv preprint arXiv:2502.07456, Feb. 2025. https://doi.org/10.48550/arXiv.2502.07456.

[11] H. Zhang, et al., "FedDUA: Federated learning with dynamic update and adaptive pruning using shared data on the server", arXiv preprint arXiv:2204.11536, Apr. 2022. https://doi.org/10.48550/arXiv.2204.11536

[12] J. Lai, et al., "pFedGPA: Diffusion-based Generative Parameter Aggregation for Personalized Federated Learning", Proc. of the AAAI Conference on Artificial Intelligence, Philadelphia, Pennsylvania, Vol. 39. No. 17, Feb. 2025. https://doi.org/10.1609/aaai.v39i17.33980.

[13] S. Han, H. Mao, and W. J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding", 4th International Conference on Learning Representations(ICLR), San Juan, Puerto Rico, May 2016. https://doi.org/10.48550/arXiv.1510.00149

[14] K. Gupta, et al., "Quantization robust federated learning for efficient inference on heterogeneous devices", Transactions on Machine Learning Research, Jun. 2022. https://doi.org/10.48550/arXiv.2206.10844.

[15] P. R. Ovi, E. Dey, N. Roy, and A. Gangopadhyay, "Mixed Quantization Enabled Federated Learning to Tackle Gradient Inversion Attacks", 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, pp. 5046-505., Jun. 2023. https://doi.org/10.1109/CVPRW59228.2023.00533.

[16] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients", Advances in Neural Information Processing Systems, Vancouver, Canada, Vol. 33, pp. 14774-14784, Dec. 2019. https://doi.org/10.48550/arXiv.1906.08935.

저자소개

임 승 호 (Seung-Ho Lim)

2001년 2월 : 한국과학기술원 전기 및 전자공학과(공학사)
2003년 2월 : 한국과학기술원 전기 및 전자공학과(공학석사)
2008년 2월 : 한국과학기술원 전기 및 전자공학과(공학박사)
2008년 3월 ~ 2010년 2월 : 삼성전자 메모리 사업부 책임연구원
2010년 3월 ~ 현재 : 한국외국어대학교 컴퓨터공학부 교수
관심분야 : 운영체제, 파일 시스템, 임베디드 시스템, DLA for AI, 비휘발성 메모리 시스템, Hadoop, HDFS, Federaged Learning, Edge Computing