# MNIST 분류를 위한 분산신경망에서 무선 자원 사용량 감소를 위한 데이터 취사선택 방식

진의환, 김종현, 이광훈, 이정섭, 김성환, 김광순
연세대학교 전기전자공학과

{jinian, jonghyun.kim, kwanghoon.lee, jungseop.lee, seonghwan, ks.kim}@yonsei.ac.kr

# Data Selection for Radio Resource Usage Reduction in Distributed Neural Networks for MNIST Classification

Eui Whan Jin, Jonghyun Kim, Kwanghoon Lee, Jungseop Lee, Seonghwan Kim, Kwang Soon Kim
School of Electrical and Electronic Engineering, Yonsei University

요 약

본 논문에서는 이동통신망 위에서 작동하는 분산신경망을 통한 협력 인공지능 실현에 필요한 데이터를 중요도에 따라 전송 여부를 결정하여 무선 자원 사용량을 줄일 수 있는 데이터 취사선택 방식을 제안하였다. 분산신경망 구동에 필요한 전송 데이터의 중요도를 판단하여 일부만 선택하여 학습 및 추론 성능을 유지하면서 무선 자원 사용량을 줄일 수 있음을 보였다.

## Ⅰ. 서 론

미래 ICT 기술은 기존 이동통신 서비스가 제공하던 사람간, 사람과 기계의 연결성을 넘어 자동화되고 지능화된 기계들에게 연결성을 제공하여 더욱 다양한 use-case 에 활용됨을 목표로 하고 있다[1]. 또한 5G 이후 다음 이동통신 시스템의 주 목적은 다수의 기계들 사이의 실시간 기계학습 및 인공지능을 통한 자율제어 기술의 인프라 제공이 될 것이라 예상된다.

이러한 기술의 실현을 위해서는 분산된 형태의 인공지능에 대한 연구가 필요하다. 이제까지는 클라우드 서버에서의 기계학습 방식이나, 비실시간 분산학습에 대한 연구가 주로 진행되고 있다[2][3]. 이러한 방법은 정보 흐름의 병목 현상과 개인 정보 보호의 어려움 및 인공지능 구현 성능 저하 등의 문제가 일어날 수 있으며, 인공 신경망에 다양한 센싱 데이터를 입력으로 받아서 구동해야하는 미래의 대규모 인공지능 인프라에서는 여러 정보를 한 곳에서 처리하는 것보다 인공신경망의 구조를 여러 기계 및 서버에 분산시켜 이동통신망을 통해 협력적으로 인공지능을 구현하는 것이 정보 이용의 효율성 측면에서 바람직하다. 분산신경망 구성에 필요한 미래 이동통신 인프라는 이동체간의 고용량 데이터를 저지연/고신뢰로 제공할 수 있는 통신 성능을 제공해야 하며, 분산신경망 사이에 전달되는 트래픽의 특성이 기존의 영상이나 음성 트래픽의 특성과는 다른 형태일 것이기 때문에 이에 대한 연구가 필요하다. 또한 이동통신 기술이 발전하고 수요가 증가함에 따라 무선데이터 트래픽이 증가하였고, 주파수 대역이 포화됨에 따라 추가 주파수 대역 확보가 어려워져 무선 자원 활용의 효율성 증대가 더욱 필요하게 되었다[4].

본 논문에서는 분산된 형태의 인공지능의 일반적인 시나리오를 차용하고, 그에 부합하는 가장 기초적인 형태의 분산신경망을 구현하여 분산신경망 구동에 필요한 전송 데이터의 특성을 파악하였으며, 전송 데이터의 특성에 따라 취사선택을 통해 데이터 전송량을 감소시켜 분산신경망의 학습 및 추론 성능의 저하 없이 무선 자원 사용량을 감소할 수 있는 방법을 제안하였다.

## Ⅱ. 본론

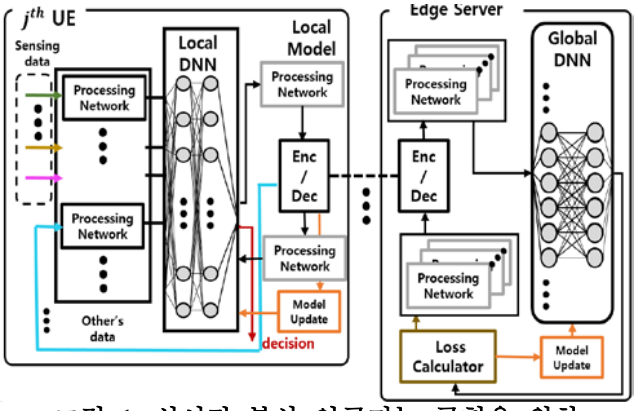


**그림 1. 실시간 분산 인공지능 구현을 위한 분산신경망 구조[5]**

분산 협력 인공지능의 일반적인 구조는 다수의 이동체와 엣지서버가 각각 인공신경망을 탑재하여 이동체들이 다양한 센싱데이터를 실시간으로 수집하여 입력 데이터로 사용하고 자체적으로 신경망을 통해 데이터를 처리한 후 이동통신망을 통해 서버로 전송해

이동체들이 보낸 데이터를 서버에서 처리하는 구조, 즉 이동체와 서버 모두 인공신경망을 통한 실시간 학습 및 협력적인 인공지능 업무를 수행하도록 그림 1 과 같이 나타낼 수 있다.

이러한 형태의 분산신경망의 예시로 이동체가 촬영한 시각 정보를 이동체에 탑재된 인공신경망이 처리하여 서버로 전송하고 이동체들이 보내온 정보를 서버가 취합하여 이동체와 서버가 협력적으로 컴퓨터 비전 태스크를 수행하는 보편적인 상황을 생각할 수 있다. 컴퓨터 비전을 위한 분산신경망의 구성을 위한 전송데이터의 특성을 파악하기 위해 가장 간단한 컴퓨터 비전 기술인 MNIST 분류 문제를 분산신경망을 통해 해결하는 방법으로 기초적인 분산 인공지능을 그림 2 과 같이 단순화시켜서 구현하였다. 이동체의 신경망은 95 개의 노드를 갖는 퍼셉트론 층 두 개로 구성되며 엣지서버의 신경망은 256 개의 노드를 갖는 다섯 개의 퍼셉트론 층으로 이루어져 있다.

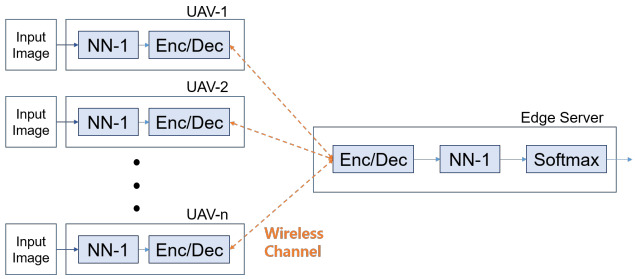


그림 2. MNIST 분류를 위한 분산신경망

분산신경망 구동을 위해 이동체와 서버 사이에는 추론 상황에서는 이동체에 탑재된 신경망의 결과값이 서버에 전달되고, 학습을 위해 일어나는 역전파 상황에서는 이동체에 탑재된 신경망의 결과값으로 추론 오차를 편미분한 값이 서버에서 이동체로 각각 이동통신망을 통해 전달된다. 추론과 학습을 위해 오가는 데이터들은 학습이 진행될수록 추론을 위한 결과값은 큰 값들만 그 숫자를 유지하고 중간값이나 작은 값들은 점점 더 작아지는 경향을 보임에 따라 추론을 위해서는 결과값들 중에서 일정 수준 이상의 큰 값들이 결정적인 역할을 할 수 있는 통계적 특성을 [6]에서 파악하였다.

| Output Value | Overall Uniform Quantization (n bits) | Quantization with Data Selection (n-1 bits) |
|---|---|---|
| $0 \le x \le median$ | $2^n$- level Uniform Quantization | $x = 0$ |
| $median \le x \le maximum * 0.99$ | | $2^n - 2$ level Uniform Quantization |
| $maximum * 0.99 \le x \le maximum$ | | $x = maximum * 0.99$ |

표 1. 분산신경망에서의 Enc/Dec 및 전송 비트수

이에 따라 이동체에 탑재된 신경망의 결과값 중에서 추론에 결정적인 역할을 하지 못하는 작은 값들은 전부 0 으로 취급하여 전송하여도 학습 및 추론 성능의 저하 없이 이동통신망을 통해 전송되는 데이터의 양을 감소시킬 수 있으며 이를 통해 통신을 위한 양자화 과정에서 양자화 비트를 줄여 통신량을 감소시킬 수 있다. 첫 번째 epoch 에서는 모든 데이터를 양자화하여 전송하고, 서버가 수신한 이동체 신경망의 결과값중 중간값을 기준값으로 결정하여 두 번째 epoch 부터는 기준값보다 작은 값들을 0 으로 치환하여 서버에 전송하였을 때 기존 양자화 비트수보다 1 비트를 줄여서 전송량을 줄일 수 있음을 표 1 에서 확인할 수 있으며, 이러한 데이터 취사선택 방식을 통해 학습 및 추론 성능의 저하가 없음을 그림 3 과 같이 나타내었다.

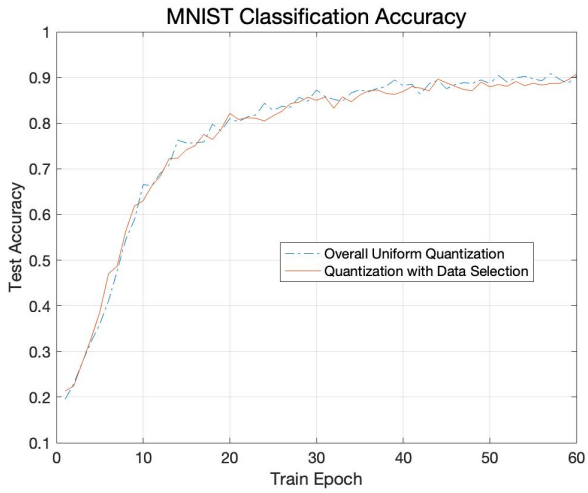


그림 3. 데이터 선택률에 따른 MNIST 분류 정확도

## Ⅲ. 결론

본 논문에서는 분산신경망 구성을 위한 전송 데이터의 통계적 특성 및 데이터 값에 따른 역할의 중요성을 파악하여 추론 상황에서 중요도가 낮은 작은 값들을 0 으로 치환하여 전송하는 방식을 통해 학습 및 추론 성능의 저하 없이 데이터 전송량을 줄여 무선 자원 효율성 증대가 가능함을 보였다.

## ACKNOWLEDGMENT

이 논문은 2019 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2019R1A2C2007982)

## 참 고 문 헌

[1] 3GPP Technical Report 38.013, “Study on Scenarios and Requirements for Next Generation Access technologies.”

[2] Rui L. Aguiar, Instituto de telecomunicacoes "White Paper for Research Beyond 5G" 20- oct-2015

[3] H. Chan et al., “Problem statement for distributed and dynamic mobility management,” internet-draft, IETF, 2011.

[4] 조용호, “5G 이동통신 표준화 동향”, 정보통신기술진흥센터, 2018. 5.

[5] 이정섭, 진의환, 김광순 “분산 학습 서비스를 위한 일반화된 아키텍처와 패킷 특성”, 2020 한국통신학회 동계종합학술발표회

[6] 진의환, 김종현, 김광순, “분산 학습 상황에서 균등 양자화를 이용한 통신 오버헤드 감소”, 2019 한국통신학회 하계종합학술발표회