선택적 상향링크 전송을 통한 통신 효율적인 보조 서버 기반 분할 학습

남혜린*, 김성륜
연세대학교

*hlnam@ramo.yonsei.ac.kr, slkim@yonsei.ac.kr

Tiny Server-assisted Communication-efficient Split Learning via Opportunistic Uplink transmission

Nam Hye Lin*, Seong-Lyun Kim
Yonsei Univ.

요 약

본 논문은 반복적인 대용량 통신을 요구하는 분산 러닝의 문제점을 해결하는 새로운 구조를 소개한다. 최근 인공지능과 통신 기술의 발전에 따라 다수의 단말이 높은 성능으로 분산된 데이터를 학습하는 분산 학습, 특히 분할 학습이 많이 연구되고 있으나, 빈번한 전송 횟수는 실제 상용되기에 어려움을 가져온다. 이에 본 연구에서는 단말에 보조 서버 모델을 탑재하여 이에 기반하여 선택적으로 서버로 은닉값을 전송함으로써 전송률을 감소시킨다. 시뮬레이션을 통해 낮은 전송률로 높은 성능을 유지하는 결과를 보였으며, 이는 실제 MEC 환경에 적용할 때 효과적일 것으로 기대한다.

Ⅰ. 서 론

최근 URLLC, mMTC 등의 서비스가 발전함에 따라 인공지능을 탑재한 단말의 수가 증대하고 있다. 이러한 고성능·고용량 어플리케이션은 대용량 데이터와 그에 따른 상당한 컴퓨팅 자원을 필요로 한다. 따라서 광범위한 지역에 분산되어 위치한 단말들의 데이터와 클라우드 컴퓨팅 서버의 자원을 활용하기 위하여 MEC (Mobile Edge Computing) 구조의 분산 러닝이 대두되고 있다. 그 중 본 논문에서는 분할 학습 (Split Learning) [1] 개념을 다룬다.

분할 학습에서는 하나의 뉴럴 네트워크가 특정 층 기준으로 분할되어 입력 쪽의 하단 부분은 복제되어 단말들이 가지고, 남은 상위 층들은 서버가 가져 최종 출력을 생성한다. 학습 시 순전파 과정에서 각 단말은 각 데이터를 분할 층까지 계산한 은닉 값(hidden activation, smashed data)을 서버에게 전송, 서버는 은닉 값을 마지막 층까지 계산하여 최종 출력을 낸다. 이를 실제 레이블 값과 비교한 손실 값에 경사하강법으로 서버를 업데이트 하고, 분할 층의 기울기를 각 단말에게 전달하여 역전파 과정을 진행한다.

이러한 분할 학습 과정의 가장 큰 문제는 통신 로드가 크다는 점이다. 학습 과정은 전체 데이터를 배치 단위로 나누어 반복하고, 이 전체는 에폭 단위로 반복되는데, 분할 층의 은닉 값과 기울기는 각각 상향 링크와 하향 링크로 매 배치마다 전송이 되어 매우 빈번한 통신이 일어난다. [2]

따라서, 본 논문에서는 순전파 시의 상향링크 전송 횟수를 줄여 통신 로드를 감소시키고, 성능은 최대한 유지하는 분할 학습 구조를 소개한다. 컴퓨팅 서버를 대신하는 **보조 서버 (Tiny server)**를 각 단말에 탑재하여, 단말에서도 출력 값을 도출할 수 있게 한다. 단말은 서버에게 은닉 값을 전송하기 전, 보조 서버를 통해 손실 값을 계산한 후, 손실 값이 커서 모델 업데이트에 큰 영향을 줄 것이라 판단할 때에만 은닉 값을 상향링크로 전달함으로써 학습 시 통신 로드를 줄인다.

Ⅱ. 본론

본 논문에서는 보조 서버의 개념을 소개한다. 단말에 위치해 서버를 대신하는 역할을 하기 때문에, 서버의 동작을 모방해야 한다. 즉, 서버가 해당 은닉 값으로 출력할 손실 값을 보조 서버가 정확히 예측해야 한다.

일반적인 MEC 환경에서 단말은 상대적으로 작은 컴퓨팅 파워를 가지기 때문에, 서버 모델보다 작은 보조 서버를 가져야 한다. 그러나 이러한 용량차에 의하여, 보조 서버는 실시간으로 학습하는 서버를 따라 동일한 방향과 속도로 변화하는 데에 어려움이 있다. 이를 최소화하기 위해 *가지치기 (pruning)*기법을 통해 두 모델 용량차를 줄이고, *지식 증류 (knowledge distillation)*를 적용하여 보조 서버를 학습시키는 두 기법을 고안한다.

1. 보조 서버 (Tiny Server) 학습 기법

가. 서버 가지치기 후 보조 서버에게 증류

지식 증류는 네트워크 깊이가 비슷할 때, 특히 선생 모델과 학생 모델의 출력 값 극명도가 비슷할 때 효과가 크다. [3]

그림 1. 보조 서버 학습 방법 (좌: '가', 우: '나')

따라서 가지치기를 통해 서버의 출력 값을 의도적으로 불분명(soft)하게 하고, 이를 보조 서버가 학습하도록 했다.

나. 서버로부터 증류 후 보조 서버 가지치기

지식 증류 시 '가'에서는 서버(선생 모델)와 보조 서버(학생 모델)의 용량 차를 극복하기 위하여 서버를 축소한 반면, 본 방법은 보조 서버의 크기를 확대한다. 큰 용량의 보조 서버를 생성, 서버와 깊이를 비슷하게 하여 증류 효과를 증대한 후, 단말의 컴퓨팅 용량에 부합하도록 가지치기를 하여 작은 보조 서버를 최종 생성한다.

2. 시뮬레이션 결과

실험은 VGG11 [4] 모델을 변형하여 합성곱(convolutional) 층 6 개와 완전연결(Fully-connected) 층 3 개로 이루어진 모델을 사용하였고, 입력 부분부터 합성곱 층 3 개까지를 단말이 갖도록 분할하였다. 단말의 개수는 4 개로, 각 단말은 아래의 식과 같이 보조 서버를 통해 예측한 서버의 손실L_{est}이 이전L_{last}에 비해 크게 줄어들지 않았을 때, 해당 은닉 값을 서버에게 상향링크 전송한다.

$$if\ not\ \ L_{last} - L_{est} < L_{last} \cdot \alpha$$

기준점 α는 0.01, 옵티마이저는 SGD 를 사용하고 학습률은 0.1 로 설정하였다.

가. 성능 및 전송률

그림 2 는 각 학습 방법에 따른 기존 전송 횟수 대비 상향링크 전송률 (Uplink Ratio), 주 모델 즉 서버의 성능 (Acc of Server), 보조 서버와 서버의 일치함 정도(Fidelity with Server)를 측정하였다.

보조 서버 없이 주 모델을 학습시키는 기본 방법, 'Without Tiny server' 와 제안하는 보조 서버를 활용한 '가, 나' 방법은 전송 횟수를 65% 정도로 줄였으나 성능은 유지하며 특히 '나'는 더 좋은 성능을 보인다.

'Without pruning'도 보조 서버를 이용하지만 '가,나'와 달리 가지치기를 제외, 지식 증류 기법만 사용한다. 이는 성능이 떨어짐을 알 수 있다. 따라서 지식 증류와 가지치기를 동시에 사용하였을 때 보조 서버는 서버의 학습 방향과 정도를 잘 따라함을 알 수 있다.

보조 서버의 출력 레이블이 서버와 일치하는 정도는 지식 증류나 가지치기를 사용하지 않는 'With Only Tiny server' 방식에 비해 제안하는 기법들이 더 높다.

나. 전송 및 컴퓨팅 로드

표 1 은 한 번의 에폭에서 발생하는 전송과 컴퓨팅 로드를 계산하였다. 제안하는 기법에서 전송률이 1 이하이므로 전송 로드에 이득이 있다. 컴퓨팅 로드에는 기존과 달리 보조 서버에 의한 계산이 추가되나, 전송률만큼 로드가 줄기 때문에 최종적으로 본 실험 환경에서는 증가하지 않는다.

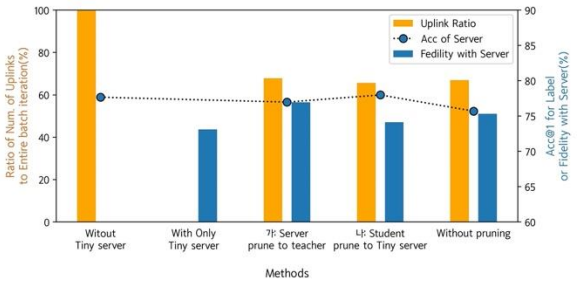

그림 2. 각 학습 방법에 따른 성능

D	데이터셋 크기	x	상향링크 전송률 ∈ [0,1]
a	은닉 값 크기	n	배치 반복 횟수
w_s, w_c, w_t	서버 모델, 단말 모델, 보조 서버 모델 크기		

	분열 학습	분열 학습 + 보조 서버
전송 로드 (상향 링크)	$\|D\|a \cdot n$	$\|D\|a \cdot x \cdot n$
전송 로드 (하향 링크)	$\|D\|a \cdot n$	$\{\|D\|a + \|w_t\|\} \cdot x \cdot n$
서버 컴퓨팅 로드	$2\|D\|\|w_s\| \cdot n$	$2 \cdot \{\|D\|\|w_s\| + \|D\|\|w_t\|\} \cdot x \cdot n$
단말 컴퓨팅 로드	$2\|D\|\|w_c\| \cdot n$	$\{(\|D\|\|w_c\| + \|D\|\|w_t\|) + \|D\|\|w_c\| \cdot x\} \cdot n$

표 1. 보조 서버 기반의 분할 학습 구조의 전송, 컴퓨팅 로드

Ⅲ. 결론

본 논문에서는 기존 분할 학습의 높은 전송 로드 문제를 해결하기 위해 보조 서버의 개념을 소개한다. 보조 서버는 단말에서 상향 링크 전송 없이 서버의 출력값과 손실을 예상할 수 있으며, 이에 따라 학습 효과가 좋다고 판단한 은닉값만 서버로 전송하여 주 모델을 학습한다. 서버를 모방하기 위하여 가지치기와 지식 증류 기반의 두가지 보조 서버 학습 방법을 제안한다. 모두 전송률은 감소하며 성능은 유지하는 결과를 보이고, 특히 큰 보조 서버를 지식 증류를 통해 학습시킨 후 단말의 컴퓨팅 자원에 부합하게 가지치기로 용량을 줄이는 방법이 더 우수하다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00347, 6G 통신을 위한 Post MAC)

참 고 문 헌

[1] S. Oh, J. Park, P. Vepakomma, S. Baek, R. Raskar, M. Bennis and S. -L. Kim, "LocFedMix-SL: Localize, Federate, and Mix for Improved Scalability, Convergence, and Latency in Split Learning,", TheWebConf 2022 (WWW conference), (2022).

[2] Singh, Abhishek, et al. "Detailed comparison of communication efficiency of split learning and federated learning." *arXiv preprint arXiv:1909.09145* (2019).

[3] Cho, Jang Hyun, and Bharath Hariharan. "On the efficacy of knowledge distillation." *Proceedings of the IEEE/CVF international conference on computer vision*. (2019).

[4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).