저데이터전송률 통신 웨어러블 디바이스를 위한 분산 뉴럴 네트워크 최적화 기법

정종훈, 이다솜, 양회석
아주대학교
{kkjjh223, ekthadl40, hyang}@ajou.ac.kr

Optimization of Distributed Neural Network on Wearable Devices with Low Data Rate Communication

JongHun Jeong, Dasom Lee, Hoeseok Yang
Ajou Univ.

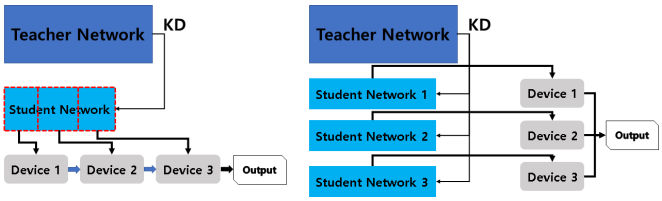
요 약

본 논문은 인체통신과 같은 저데이터전송률 통신을 활용하는 웨어러블 디바이스에서 효율적으로 분산 뉴럴 네트워크를 구동하기 위한 NoNN(Network of Neural Network)기법을 소개하고, 통신 실패에 따른 추론 정확도 열화가 크다는 NoNN의 단점을 개선한다. 제안하는 분산 뉴럴 네트워크는 생성과정에서 지식 추출의 결과를 중복적으로 분산 디바이스에 반영함으로써 일부 디바이스에서 통신 실패가 발생하여도 최종 출력 결과의 변화가 적어 통신 실패에 강인하다. 8개의 디바이스로 뉴럴 네트워크를 분산시켜 기존 NoNN과 제안한 기법을 비교실험한 결과, 3개의 디바이스에서 통신 실패가 발생할 때 평균 1.164%, 최악의 경우 11.22% 개선되었고 7개의 디바이스에서 통신 실패가 발생할 때 평균 18.741% 최악의 경우 54.22% 개선되어 제안하는 분산 뉴럴 네트워크가 통신 실패에 강인한 것을 검증하였다.

Ⅰ. 서 론

최근 다양한 사물인터넷(Internet of Things) 장치의 확산과 함께 인체에 착용하는 다양한 웨어러블 디바이스가 개발되고 있다. 또한, 최근 인공지능 기술의 발전과 더불어 이러한 웨어러블 디바이스 상에서도 뉴럴 네트워크(Neural Network)기반 응용프로그램을 구동할 필요성이 대두되고 있다. 일반적인 내장형 시스템에서는 메모리 용량 등 자원제약이 심하므로 단일 디바이스만으로는 뉴럴 네트워크 구동이 어렵다. 하지만 통신장치가 부착된 웨어러블 디바이스에서는 뉴럴 네트워크를 다수의 디바이스에 분산 하여 병렬적으로 구동할 수 있다.

일반적으로 웨어러블 디바이스들은 블루투스나 와이파이와 같은 근거리 통신 프로토콜을 사용하여 통신하나 군용 제품이나 재난 환경 등의 특수한 상황에서는 사람의 몸을 매질로 하는 인체 통신을 사용하여 통신할 것으로 예측된다. 이러한 인체통신은 추가적인 기존 유무선 통신에 비하여 높은 편리성과 용이성을 가지나 낮은 데이터 전송률을 가지는 단점이 있다.[1] 인체 통신과 같이 데이터 전송률이 낮고 불안정한 통신에 의존하는 경우, 웨어러블 디바이스에서 사용할 분산 뉴럴 네트워크 기법은 통신량이 적고 일부 디바이스에서 통신에 실패하여도 결과를 도출할 수 있도록 통신 실패에 강인해야 한다.



(a) Distributed NN (b) NoNN

그림 1 분산 뉴럴 네트워크 생성 과정: (a) 기존 기법 및 (b) NoNN

그림 1(a)은 기존의 일반적인 분산 뉴럴 네트워크 생성 과정을 도식화한 것이다. 즉, 먼저 지식추출(Knowledge Distillation, KD)을 통하여 메모리 요구량과 연산량이 큰 고성능(teacher) 네트워크의 최종결과를 학습한 서브(student) 네트워크 한 개를 생성한다. 이 student 네트워크도 단일 디바이스에서 구동하기에 여전히 크기 때문에 레이어 단위로 여러 디바이스에 분산 시킨다. 그러나 이 경우 레이어 간 전달 데이터의 양이 커서 통신량과 통신 빈도가 높을 뿐 아니라, 하나의 추론 결과를 얻기 위해서 모든 디바이스의 결과가 필요하므로 하나의 디바이스에서라도 통신이 실패할 경우 결론 도출이 불가능하다는 단점이 있다.

이러한 단점을 보완하기 위해 그림 1(b)와 같은 NoNN(Network of Neural Network)[2]이라는 분산 뉴럴 네트워크 기법이 제안되었다. NoNN의 student 네트워크들은 모두 독립적인 작은 네트워크이며, 이 독립적인 네트워크들 각각의 최종 결과들을 concatenate 하여 최종 결과를 도출한다. 이 경우, student network들은 각각의 독립적인 네트워크로 독자적으로 결과 도출이 가능하지만, 상대적으로 별개 네트워크의 정확도는 많이 떨어져 통신 실패의 경우 추론 정확도 열화가 크다. 본 논문에서는 이러한 단점을 해결하기 위한 개선된 분산 뉴럴 네트워크 최적화 기법을 제안한다.

Ⅱ. 배경 지식

NoNN 기법에서는 teacher 네트워크의 필터(Knowledge Distillation을 통해 추출된 지식)들을 한쪽에 치우치지 않도록 Activation Hub(AH) 규칙에 의거하여 필터를 여러 개의 group으로 나눈다. AH 규칙은 동일한 종류의 입력에 대하여 큰 중요도를 가지는 필터들을 식(1)을 기준으로 나눈다. a_i는 teacher 네트워크의 마지막 layer에서 i번째 필터의 출력 평균, a_j는 j번째 필터의 출력 평균을 의미할 때 AH 규칙은 수식(1)로 표현된다.

$$AH(i,j) = \sum a_i a_j |a_i - a_j| \quad (1)$$

분산 네트워크의 효율적인 분할을 위하여 각 필터들을 정점(vertex)으로 하고 그 필터 간의 간선(edge)들의 가중치를 AH값으로 가지는 그래프 형태의 자료구조를 생성할 수 있다. 이를 소셜 네트워크 등을 분석하는 네트워크 사이언스에서 활용되는 분할 알고리즘[3]을 활용하여 별개의 그룹으

로 분리할 수 있는데, 이를 community structure이라고 한다. 이렇게 teacher 네트워크의 각 필터들을 다수 개의 그룹으로 나눈다. 이 때 동일한 입력에 대해 중요도가 높아 출력 값이 큰 필터들은 AH 값이 작아 다른 group으로 나뉘어 분산된다. AH 규칙으로 나뉜 group들은 네트워크 마다 크기와 개수가 다르기 때문에 NoNN은 group들을 합쳐 분산할 네트워크 수와 같은 필터들의 집합체로 구성하여 생성한다. 이 때 생성된 필터들의 집합체를 part라 칭하며 각 part들은 모두 비슷한 수의 필터로 구성되도록 생성한다. 그 후, 생성된 part들은 하나의 서브(student) 네트워크를 훈련 시킨다. 이러한 NoNN의 접근법은 통신에 실패한 경우 teacher 네트워크의 일부 필터를 사용할 수 없으므로 추론 정확도 열화가 크다.

그림 2는 NoNN에서 4개의 student 네트워크를 생성하는 예시이다. 이 경우 5개의 필터 그룹에서 4개의 디바이스로 분산되는 뉴럴 네트워크를 구성하여야 하므로, 4개의 part를 생성한다. 이렇게 생성된 NoNN 네트워크에서는 student 1에서 통신이 실패할 경우 part1을 사용할 수 없고 final feature의 0~9번 요소는 0 값을 가지므로 큰 추론 정확도 열화를 보인다. 이 문제점을 해결하기 위하여 본 논문에서는 통신의 실패로 인하여 일부 student 네트워크의 결과가 전달되지 않더라도 추론 정확도를 유지할 수 있는, 통신 실패에 강인한 분산 뉴럴 네트워크 최적화 기법을 제안한다.

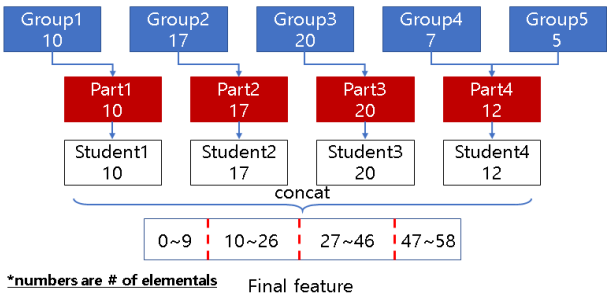


그림 2 5개의 group, 4개의 part의 경우 NoNN 생성 예시

Ⅲ. 제안하는 분산 뉴럴 네트워크 생성 기법

그림 3은 제안하는 분산 뉴럴 네트워크 생성 기법이 그림 2의 예에 대하여 어떻게 동작하는지 예시한다. 동일한 AH 기법을 통해 도출된 다섯 개의 필터 group을 합쳐 2개의 part를 만든 뒤 각 part 별로 2개의 student를 생성한다. 그 뒤 각 student에서 결과를 전송하게 되면 같은 part에 의해 생성된 student의 결과는 평균을 내서 2개의 part의 결과를 concatenate 하게 된다. 제안하는 분산 뉴럴 네트워크는 part의 개수를 디바이스 수보다 적게 구성한 후 디바이스에 각 part의 지식이 중첩되게

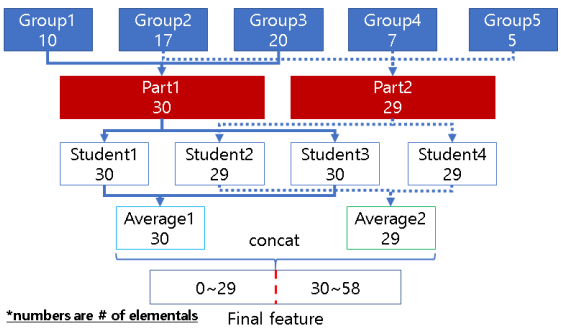


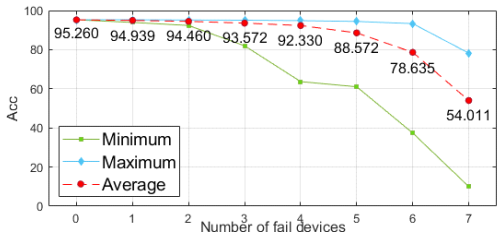
그림 3 제안하는 기법을 활용한 분산 뉴럴 네트워크 생성 예시

반영하여 student1에서 통신 오류가 발생해도 part1을 사용할 수 있으며 평균값을 이용하기 때문에 final feature map에는 큰 변화가 없어 통신 실패로 인한 추론 정확도 열화가 개선된다.

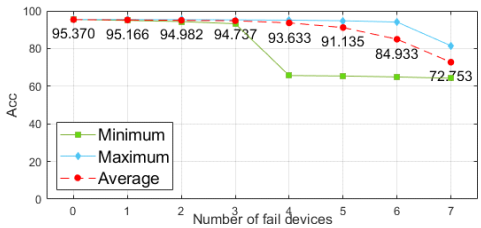
Ⅳ. 실험 결과

제안하는 분산 뉴럴 네트워크의 성능 평가를 위해 8개의 student 네트워크를 가지도록 NoNN과 제안하는 분산 뉴럴 네트워크를 학습시켰다. 이 때, teacher 네트워크는 WRN40-4를 이용한 cifar10 dataset 이미지 분류 네트워크를 사용하였고 student는 NoNN논문에서의 실험과 동일한 형태의 WRN을 사용하여 훈련시켰다. 학습된 네트워크들은 일부 student 네트워크의 통신 실패를 가정하여 출력 값을 0으로 설정하는 실험을 구성하였다. 통신 실패를 가정한 실험은 8개의 student에서 일어날 수 있는 모든 경우의 수에 대해 실험을 진행하였다.

실험 결과, 그림 4와 같이 통신 실패가 3개 발생하는 경우 정확도가 평균 1.164%, 최악의 경우 11.22% 개선되었다. 이는 통신 실패의 개수가 커지는 경우 더 뚜렷하게 나타나며 통신 실패가 7개 나타나는 경우 평균 18.741%, 최악의 경우 54.22% 개선을 보여 제안하는 분산 뉴럴 네트워크가 통신 실패에 강인한 것을 확인할 수 있었다.



a. NoNN



b. 제안하는 네트워크

그림 4 cifar10 이미지 분류 분산 네트워크 통신 실패 실험 결과

Ⅴ. 결론

본 논문에서는 낮은 데이터전송률 통신을 활용한 웨어러블 디바이스를 위한 분산 뉴럴 네트워크 최적화 기법에 대해 소개하였다. 제안하는 기법은 NoNN의 분산 뉴럴 네트워크 생성과정을 변경하여 필터들의 묶음(part) 갯수를 디바이스 수보다 적게 구성하고, 여러 part들이 단일 디바이스에 중복 반영되게 함으로써 통신 실패에 강인하도록 개선하였다. 실험 결과 제안하는 방식이 기존 NoNN보다 개별 디바이스들의 통신 실패에 강인한 것을 확인할 수 있다. 통신 실패의 개수가 커지는 경우 제안하는 네트워크의 추론성능 개선이 더 뚜렷하게 나타나 7개의 디바이스에서 통신 실패가 발생하는 경우 더 큰 개선을 보여 제안하는 분산 뉴럴 네트워크가 통신 실패에 강인한 것을 검증하였다.

ACKNOWLEDGMENT

본 연구는 방위사업청과 국방과학연구소가 지원하는 미래전투체계 네트워크기술 특화연구센터 사업의 일환으로 수행되었습니다.(UD190033ED)

참 고 문 헌

[1] 강성원; 박형일; 박경환. 인체통신 기술 현황 및 전망. 2013.

[2] Bhardwaj, Kartikeya, et al. "Memory-and communication-aware model compression for distributed deep learning inference on iot." *ACM Transactions on Embedded Computing Systems* (TECS) 18.5s (2019): 1-22.

[3] Newman, Mark EJ. "Modularity and community structure in networks." *Proceedings of the national academy of sciences* 103.23 (2006): 8577-8582.