오프로딩 환경을 위한 신경망 전송-연산 연동 기법

빈경민, 김세연, 이경한
서울대학교 전기 정보 공학부 뉴미디어통신공동연구소

kmbin@snu.ac.kr, seyeon625@snu.ac.kr kyunghanlee@snu.ac.kr

Pipelined Transmission and Computation for Offloaded Neural Network Inference

Kyungmin Bin, Seyeon Kim, Kyunghan Lee
Department of Electrical and Computer Engineering and INMC
Seoul National University

요 약

신경망 연산은 많은 컴퓨팅 자원을 필요로 하기 때문에 제한적인 컴퓨팅 자원 및 쿨링 능력을 가진 모바일 기기에서 직접적인 신경망 추론 연산을 하기에는 응답 지연 시간 및 에너지 소모량 측면에서 매우 비효율적이다. 이로 인해 클라우드 서버에 신경망 연산을 요청하는 신경망 오프로딩 기법이 활발히 연구되고 있다. 하지만, 기존의 신경망 오프로딩 기법들은 데이터 수신의 보장이 필요한 순차적 전송 연산 방식이므로 실시간으로 변화하는 네트워크 상황에서 응용에서 필요로 하는 응답 지연 시간을 달성하기 매우 어렵다. 본 논문에서는 오프로딩 환경에서 신경망 연산의 응답 지연 시간 단축을 위한 신경망 연산-전송 기법을 제시한다. 제안하는 기법은 데이터 수신의 보장이 불필요하여 신경망 연산의 조기 시작이 가능하여 기존의 순차적 연산 기법 대비 연산 종료 시점을 앞당겨 응답 지연 시간을 매우 단축시킬 수 있다.

Ⅰ. 서 론

심층 신경망 (DNN) 추론 연산은 많은 컴퓨팅 자원을 필요로 하므로 제한적인 컴퓨팅 자원 및 쿨링 능력을 가진 모바일 기기에서 직접적인 신경망 추론 연산을 하기에는 응답 지연 시간 및 에너지 소모량 측면에서 비효율적이다. 이로 인해 모바일 기기에서의 직접적인 연산 보다는 컴퓨팅 자원이 풍부하고 쿨링 상황이 충분히 갖추어진 클라우드 서버에 신경망 연산을 요청하는 신경망 오프로딩 기법이 많이 선호되고 있다 [1, 2]. 기존의 오프로딩 기법에 대한 연구는 모바일 기기에서 서버로 데이터의 전송 및 수신이 완료된 후 신경망 연산을 진행하는 순차적 방식으로 연구가 진행되어 왔다. 이러한 순차적 방식은 신경망 연산에 필요한 모든 데이터 수신이 보장되어야 연산의 시작이 가능하다는 점에 있어 실시간으로 변화하는 네트워크 상황에서 응용에서 필요로 하는 응답 지연 시간을 달성하기 매우 어렵다. 본 논문에서는 응답 지연 시간 단축을 위해 오프로딩 환경에서 신경망 전송-연산 연동 기법을 제안한다. 제안하는 기법을 통해 기존의 모든 데이터 수신 보장이 필요한 방식과는 달리, 신경망 연산 시작을 위한 모든 데이터 수신의 보장이 불필요하여 연산의 조기 시작이 가능하므로 연산의 종료 시점을 매우 앞당겨 응답 지연 시간을 단축시킬 수 있다.

Ⅱ. 본론

신경망 추론 연산은 연속된 신경망 계층의 연산으로 이루어져 있다. 일반적으로 신경망 계층 연산은 입력 데이터 행렬과 신경망 가중치 (Weight) 행렬의 행렬 곱 연산으로 이루어져 있다. 하지만 여기서 주목해야 하는 점은, 신경망 연산의 특성 상 한 신경망 계층의 연산은 한 번의 행렬 곱 연산으로 이루어져 있는 것이 아니라 다중 행렬 곱 연산으로 이루어져 있다는 점이다. 예를 들어, 컨볼루션 (Convolution) 연산의 경우, 일반적으로 가중치 행렬의 크기는 입력 행렬의 크기보다 매우 작으므로 (3x3 등) 매우 많은 입력 행렬과 가중치 행렬의 곱을 통해 출력 행렬을 연산하게 된다. 이 때, 한 번의 행렬 곱으로 인해 출력 행렬의 일부분을 연산하게 되고, 출력 행렬은 다음 신경망 계층의 입력 행렬로 사용되어 같은 과정을 통해 다음 신경망 계층의 입력 행렬을 연산해내게 된다. 신경망 연산은 이러한 신경망 계층 연산들의 행렬 곱 연산들이 서로 의존성을 갖게 된다. 제안하는 기법에서는 이러한 신경망 연산의 특성에 착안하여 신경망 계층 연산의 하나의 행렬 곱을 신경망 연산의 기초 연산 단위로 정의한다. 신경망 계층 연산들을 기초 연산 단위로 분해 및 분석하여 신경망 연산의 의존성 그래프를 만들 수 있다.

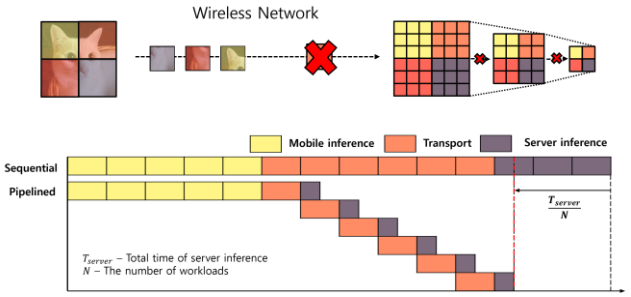

그림 1. 신경망 전송-연산 연동 기법 개념도

그림 1 은 이러한 신경망 의존성 그래프를 활용한 신경망 전송-연산 연동 기법의 개념도를 나타낸다. 신경망 연산 의존성 그래프에 따라 입력 데이터의 일부만 수신하게 된다면 그림 1 의 상단처럼 입력 값의 다른 부분들이 수신되지 않더라도 신경망 연산의 시작이 가능하게 된다. 이를 통해 그림 1 의 하단처럼 기존의 데이터 수신의 보장이 필요한 순차적 전송-연산 방식에 대비해 연산의 종료 시점을 매우 앞당길 수 있다.

Ⅲ. 결론

본 논문에서는 오프로딩 환경에서 기존의 순차적 전송-연산 방식에 대비해 연산의 종료 시점을 매우 앞당길 수 있는 신경망 연산의 응답 지연 시간 단축을 위한 신경망 전송-연산 연동 기법을 제시한다. 제안하는 기법에서 신경망 연산의 기반이 되는 행렬 곱 연산의 원리를 분석하여 기초 연산 단위를 정의하고 이를 기반으로 한 의존성 그래프를 통한 전송-연산의 연동 기법을 제시한다.

ACKNOWLEDGMENT

이 연구는 2022 년도 과학기술정보통신부의 재원으로

한국연구재단의 지원을 받아 수행되었음

(No.2022R1A5A1027646).

참 고 문 헌

[1] Kang, Yiping, et al. "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge." *ACM SIGARCH Computer Architecture News* 45.1 (2017): 615-629.

[2] Laskaridis, Stefanos, et al. "SPINN: synergistic progressive inference of neural networks over device and cloud." *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 2020.