이동 통신 환경에서 모바일 신경망 추론 연산 오프로딩을 위한 신경망 계층 파이프라인에 대한 연구

빈경민, 이경한
서울대학교 전기 정보 공학부 뉴미디어통신공동연구소

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

Pipelined DNN Layer for Offloading in Cellular Networks

Kyungmin Bin, 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 자원을 가진 모바일 기기에서 직접적인 신경망 추론 연산을 하기에는 어려움이 있다. 이러한 문제점을 해결하기 위해 오프로딩을 통한 신경망 분할 추론 연산 기법들이 많이 제시되고 있다. 하지만, 신경망 추론의 중간 데이터의 큰 크기와 이동 통신 환경에서 제한적인 상향링크 트래픽 대역폭으로 인하여 오프로딩 시 긴 데이터 전송 시간을 겪게 된다. 본 논문은 신경망 분할 추론을 위한 오프로딩시 긴 데이터 전송 시간으로 인한 연산 시간 저하를 해결하기 위한 신경망 계층 연산의 파이프라인 기법을 제시한다.

Ⅰ. 서 론

심층 신경망 (DNN) 추론 연산은 많은 컴퓨팅 자원을 필요로 하기 때문에 제한적인 컴퓨팅 자원을 가진 모바일 기기에서 직접적인 신경망 추론 연산을 하기에는 어려움이 있다. 이러한 문제점을 해결하기 위해 신경망의 상위 계층은 모바일에서 진행하고 중간 연산 결과물을 서버로 전송한 뒤 서버에서 하위 계층을 처리하고 결과를 돌려받는 방식의 신경망 분할 연산 기법에 대한 연구가 활발히 진행되고 있다 [1, 2]. 하지만, 이동 통신 환경에서는 상향링크 대역폭이 제한적이기 때문에 큰 크기의 중간 데이터를 전송하는 데에 긴 시간이 필요하다. 따라서, 이동 통신 환경에서 긴 데이터 전송 시간을 보완할 수 있는 새로운 신경망 연산 추론 기법이 필요하다. 본 논문에서는 이를 위해 신경망 계층 연산 파이프라인 기법을 제안한다. 파이프라인을 통해 기존의 순차적 방식과는 달리, 서버에서는 모든 데이터를 수신한 뒤 신경망 계층 연산을 진행하지 않고 일부 데이터만 수신한 뒤 바로 연산이 가능하게 되어 긴 데이터 전송 시간을 보완할 수 있다. 본 논문에서는 신경망 분할 추론을 위한 신경망 계층 연산 파이프라인 기법을 제시한다.

Ⅱ. 본론

신경망 계층 추론 연산 파이프라인에 있어서 가장 중요한 것은 신경망 계층 추론 연산을 효과적으로 분할하는 것이다. 다양한 신경망 계층 중에서도 가장 널리 이용되는 것은 컨볼루션 (Convolution) 계층이다. 컨볼루션 계층의 추론 연산은 $k \times k \times c$ 의 크기를 가진 커널이 입력 텐서 위를 움직이면서 반복적인 컨볼루션 연산을 통해 진행된다. 여기서, k 는 커널의 가로와 세로 크기 c 는 입력 이미지의 채널의 개수다. 따라서, 입력 이미지를 채널 방향으로 분할하여 데이터를 전송하면 하나의 컨볼루션 연산을 단위로 컨볼루션 계층 연산을 분할하여 파이프라인이 가능하다.

본 논문에서 제안하는 시스템 구조는 그림 1 과 같다. 주황색 박스는 하나의 데이터를 나타낸다. i 번째 계층에서 모바일로부터 데이터를 받고 Chunk creator 가 한 번의 연산을 진행할 수 있을 만큼의 데이터가 큐에

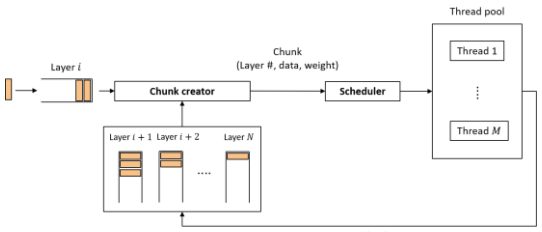

그림 1. 신경망 계층 추론 연산 파이프라인 구조

쌓였을 때 하나의 연산 청크를 만든 뒤 스케줄러에 보낸다. 스케줄러는 각각의 쓰레드에 청크를 보내고 쓰레드는 그 결과를 해당하는 다음 계층의 큐에 보낸다. 이 과정을 최종 결과물이 나올 때까지 진행을 한다. 이러한 신경망 계층 연산 파이프라인을 통해 전송과 연산 및 신경망 계층간의 파이프라인이 가능해지며 이동 통신 상황에서 긴 데이터 전송 시간을 보완할 수 있게 된다.

Ⅲ. 결론

이동 통신 환경에서 신경망 분할 추론을 위한 오프로딩시 큰 데이터 크기와 제한된 상향링크 대역폭으로 인하여 긴 데이터 전송 시간을 겪게 된다. 따라서, 본 논문에서는 신경망 분할 추론을 위한 신경망 계층 연산의 파이프라인 기법을 제시한다.

ACKNOWLEDGMENT

이 논문은 2021 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

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.