https://doi.org/10.7840/kics.2023.48.4.405

신경망 분할 기반 인공신경망 추론의 에너지 효율 개선에 관한 연구

윤 상 석◆°

Energy-Efficient Artificial Neural Network Inference Based on Neural Network Partitioning

Sangseok Yun◆°

요 약

본 논문에서는 인공신경망 추론 과정의 에너지 효율에 관한 연구를 수행한다. 특히, 신경망 분할 기법을 활용해 추론 과정의 에너지 효율을 개선하는 협력 추론 기법을 제안한다. 제안 기법은 먼저 인공 신경망을 분할하고 무선 통신을 활용해 분할한 인공 신경망을 모바일 단말 및 에지 서버에서 협력적으로 추론함으로써 에너지 효율을 개선한다. 모의 실험을 통해 제안 기법이 인공 신경망 추론의 에너지 효율을 개선할 수 있음을 확인하였다.

Key Words : Artificial neural network, energy efficiency, cooperative inference

ABSTRACT

In this paper, we investigate an energy-efficient inference of an artificial neural network. In particular, we propose an energy-efficient artificial neural network inference process based on neural network partitioning, in which the artificial neural network is partitioned and distributed between a mobile end device and an edge server. Then, by performing cooperative inference via wireless communications, the amount of energy consumption required for the entire inference process can be reduced. Simulation results show that the proposed scheme can improve the energy efficiency of practical inference tasks.

Ⅰ. 서 론

심층 학습 기술의 발전으로 인해 인공지능 알고리즘이 다양한 분야에서 높은 성능을 보이며 성공적으로 적용되고 있다. 하지만, 일반적으로 인공지능 알고리즘이 높은 성능을 획득하기 위해서는 인공지능 알고리즘의 인공 신경망을 구성하는 은닉 계층 (hidden layer)의 수 및 각 은닉 계층을 이루는 은닉 노드 (hidden node)의 수가 충분히 많아야 한다. 이는 인공지능 알고리즘의 연산에 소요되는 연산량 및 소비 전력의 급격한 증가를 유발하고[1], 때문에 IoT (Internet of Things) 단말 등 비교적 낮은 연산 능력과 작은 배터리를 보유한 모바일 단말에서 고성능 인공지능 알고리즘의 활용을 어렵게 한다.

최근, 신경망 분할을 통해 인공 신경망 추론에 소요되는 지연 시간을 저감하는 연구가 다수 수행되었으나[2-4], 인공 신경망 추론에 소요되는 소비 전력에 관한 연구는 부족한 실정이다[5]. 이에 본 연구에서는 신경망 분할 기반 인공 신경망 추론 에너지 효율 개선 기법에 관한 연구를 수행한다. 특히, 본 논문에서는 인공 신경망을 분할하고, 무선 통신을 활용해 분할한 인공 신경망을 단말과 고성능 에지 서버가 협력적으로 추론함으로써 인공지능 알고리즘에 소요되는 에너지 소모를 저감하는 기법을 제안한다. 제안하는 기법은 연산 에너지 효율, 통신 에너지 효율 등 다양한 시스템 매개변수에 따라 신경망 분할 지점을 적응적으로 선택함으로써 인공지능 알고리즘의 에너지 효율을 극대화할 수 있음을 모의 실험을 통해 확인하였다.

Ⅱ. 시스템 모델

본 논문에서는 한정된 배터리를 가진 모바일 단말이 L개의 연속된 블록 (혹은 레이어)으로 구성된 인공 신경망[3]의 추론 연산을 수행하는 시스템 모델을 고려한다. 특히, 인공 신경망의 l번째 블록의 추론을 위해서는 N_l 회의 부동소수점 연산 (FLOPs, floating point operations)이 필요하고, l번째 블록의 출력 값 (feature

※ 이 논문은 부경대학교 자율창의학술연구비(2021년)에 의하여 연구되었음.
◆° First and Corresponding Author : (ORCID:0000-0002-7961-1394) Pukyong National University Department of Information and Communications Engineering, ssyun@pknu.ac.kr, 조교수, 정회원
논문번호 : 202301-01-011-A-LU, Received January 25, 2023; Revised February 13, 2023; Accepted February 13, 2023

vector)을 나타내는데 필요한 비트의 수는 P_l인 인공 신경망 추론 시나리오를 고려한다.

모바일 단말에 탑재된 프로세서의 연산 에너지 효율 (1 [J]의 에너지 당 연산 가능한 부동소수점 연산의 횟수)을 E_d [FLOPs/J]라 하면, 모바일 단말이 인공 신경망의 l번째 블록의 추론 연산을 수행하기 위해 필요한 에너지는 N_l/E_d가 된다. 결과적으로 인공 신경망 전체의 추론을 위해 필요한 에너지는 $\sum_{l=1}^{L} N_l/E_d$ 가 되고, 모바일 단말이 보유한 배터리의 용량을 B [J]라 하면 단말은 최대 $BE_d/\sum_{l=1}^{L} N_l$ 회의 인공 신경망 추론을 수행할 수 있다.

본 논문에서는 인공 신경망 추론을 위해 소요되는 에너지를 저감하고 모바일 단말의 추론 에너지 효율을 개선하기 위한 신경망 분할 기반 협력 추론 기법을 제안한다.

Ⅲ. 협력 추론 기법

본 논문에서 제안하는 협력 추론 기법은 신경망 분할 및 에지 서버와의 협력을 활용한다. 제안하는 기법은 다음과 같이 4단계로 나누어 동작한다: ① 단말은 인공 신경망의 첫번째 블록부터 분할 지점 (예를 들어, ℓ 번째 블록)까지의 연산을 수행한다; ② 단말은 중간 결과, 즉 ℓ 번째 블록의 출력을 무선 통신을 활용해 에지 서버에 전송한다; ③ 에지 서버는 단말로부터 수신한 중간 결과를 (ℓ + 1) 번째 블록의 입력으로 사용해 나머지 블록, 즉 (ℓ + 1) 번째 블록부터 L 번째 블록까지의 연산을 수행한다; ④ 마지막으로 에지 서버는 최종 추론 결과를 무선 통신을 활용해 단말에 피드백 한다.

이처럼 제안하는 협력 추론 기법을 사용하는 경우 단말은 인공 신경망 전체를 추론하는 대신 ① 단계, 즉 첫번째 블록부터 ℓ ($\leq L$)번째 블록까지의 연산만을 수행한다. 따라서 단말에서 요구되는 추론 연산의 에너지 소모량은 $\sum_{l=1}^{L} N_l/E_d$에서 $\sum_{l=1}^{\ell} N_l/E_d$로 감소할 수 있다. 하지만 제안하는 협력 추론 기법을 사용하는 경우 ③ 단계 에지 서버에서 발생하는 연산 에너지 소모 뿐만 아니라, ②, ④ 단계에서 단말과 에지 서버가 필수적으로 수행해야 하는 무선통신으로 인한 추가적인 통신 에너지 소모가 발생하게 된다. 즉, 분할 지점이 ℓ 번째 블록인 협력 추론에서 발생하는 종단 간 에너지 소모량은 ①, ③ 단계의 연산 에너지 소모량과 ②, ④ 단계의 통신 에너지 소모량의 합으로 나타낼 수 있다.

본 논문에서는 고성능 프로세서의 원활한 운용을 위해 전원에 연결되어 충분한 에너지를 공급받는 고성능 에지 서버를 고려한다[6]. 때문에 에지 서버는 에너지 사용량에 민감하지 않으며, 따라서 협력 추론의 에너지 효율에 에지 서버, 즉 ③단계에서의 연산 에너지 소모량 및 ④단계에서의 통신 에너지 소모량은 고려하지 않는다. 반면, 모바일 단말의 경우 한정된 배터리를 이용하여 구동되기 때문에 에너지 사용량에 민감하므로 협력 추론의 에너지 효율에 모바일 단말에서의 에너지 소모량, 즉 ①단계에서의 연산 에너지 소모량 및 ②단계에서의 통신 에너지 소모량만을 고려한다.

단말의 통신 에너지 효율 (1 [J]의 에너지 당 전송 가능한 비트의 수)을 E_c [bits/J]라 하면, 분할 지점이 ℓ 번째 블록이고 단말의 연산 및 통신 에너지 효율이 각각 E_d, E_c인 협력 추론의 종단 간 에너지 소모량 $\mathcal{E}(\ell|E_d, E_c)$은 아래 수식 (1)과 같이 나타낼 수 있다.

$$\mathcal{E}(\ell|E_d, E_c) = \sum_{l=1}^{\ell} \frac{N_l}{E_d} + \frac{P_\ell}{E_c}. \quad (1)$$

결과적으로 연산 및 통신 에너지 효율에 따라 협력 추론의 종단 간 에너지 소모를 최소화하는 최적의 분할 지점 ℓ^*는 수식 (2)에 나타낸 최적화 문제를 해결함으로써 찾을 수 있다.

$$\ell^* = \underset{\ell \in \{1,2,\ldots,L\}}{\operatorname{argmin}} \mathcal{E}(\ell|E_d, E_c). \quad (2)$$

따라서 제안하는 신경망 분할 기반 협력 추론 기법을 활용했을 때의 종단 간 에너지 소모량은 $\mathcal{E}(\ell^*|E_d, E_c)$가 되며, 제안하는 협력 추론 기법을 활용했을 때 단말은 최대 $B/\mathcal{E}(\ell^*|E_d, E_c)$ 회의 인공 신경망 추론을 수행할 수 있게 된다.

Ⅳ. 모의 실험

이 장에서는 제안한 신경망 분할 기반 협력 추론 기법의 성능을 확인하기 위해 이미지 분류 분야의 심층 학습에서 널리 사용되는 MobileNetV2 인공 신경망[7]을 사용해 모의 실험을 수행하였다. MobileNetV2 인공 신경망을 최소 분할 단위인 residual block 단위로 구분하면 L = 20개의 블록으로 구분될 수 있다. 이 때, 각 블록의 출력 값을 나타내는데 필요한 비트의 수는 각 블록의 출력 이미지의 픽셀 수와 데이터 타입

에 따른 픽셀 당 비트 수의 곱을 이용해 계산할 수 있다. 본 논문에서는 이미지의 데이터 타입이 단정밀도 부동 소수점 (single-precision floating point)인 경우를 고려하였으며, 따라서 하나의 이미지 픽셀을 표현하는데 32비트를 사용한다. 한편 MobileNetV2 인공신경망을 구성하는 각 블록의 추론을 위한 부동 소수점 연산의 횟수 (FLOPs)를 Google의 인공지능 라이브러리인 TensorFlow의 프로파일링 API를 이용해 측정하였다. 각 블록의 추론을 위한 부동 소수점 연산의 횟수를 각 블록의 출력 값을 나타내는데 필요한 비트의 수와 함께 그림 1에 도시하였다.

그림 1에서 각 블록의 추론을 위한 부동소수점 연산의 수 N_l은 블록별로 비교적 유사하게 유지되는데 반해 각 블록의 출력을 나타내는데 필요한 비트의 수 P_l은 후반부 블록으로 갈수록 감소하는 경향을 확인할 수 있다. 따라서, 분할 지점을 인공 신경망의 초반부 블록으로 선택하는 경우 단말이 추론 연산에 소모하는 에너지는 저감할 수 있지만 많은 수의 데이터 비트를 무선통신을 이용해 전송해야 하므로 통신에 많은 에너지를 소모하게 되고, 반대로 분할 지점을 인공 신경망의 후반부 블록으로 선택하는 경우 통신에 소요되는 에너지는 저감할 수 있지만 추론 연산에 많은 에너지를 소모하게 된다. 이처럼 인공 신경망 분할 지점 ℓ이 종단 간 에너지 소모량의 증가와 감소 사이의 tradeoff를 제어하므로 연산 에너지 효율, 통신 에너지 효율 등 시스템 매개변수에 따라 최적의 분할 지점이 존재할 것으로 추측되며, 최적의 분할 지점을 활용해 협력 추론을 수행할 때 에너지 효율을 개선할 수 있을 것이라 예상할 수 있다.

다음으로 그림 2에 분할 지점에 따른 종단 간 에너지 소모량을 도시하였다. 해당 실험에서 사용한 연산 에너지 효율 매개변수는 대표적인 저전력 연산 장치인 Raspberry Pi에서 CPU를 풀 로드했을 때의 연산

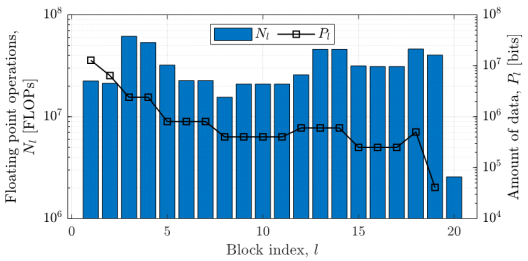

그림 1. 블록 별 연산의 수 N_l [FLOPs] 및 출력의 크기 P_l [bits]
Fig. 1. The block-wise values of N_l [FLOPs] and P_l [bits]

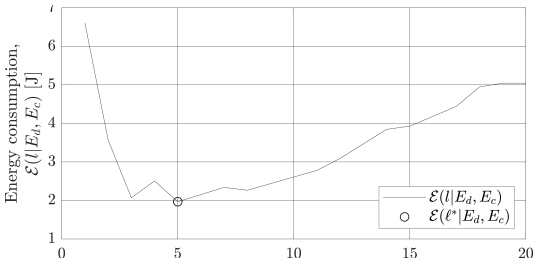

그림 2. 신경망 분할 지점 별 종단 간 에너지 소모량 및 최적의 분할 지점
Fig. 2. End-to-end energy consumption versus partitioning point and the optimal partitioning point

에너지 효율 $E_d = 1.22 \times 10^8$ [FLOPs/J]이며[8], 통신 에너지 효율 매개변수는 동일한 Raspberry Pi의 Wi-fi 업로드 환경에서 데이터 전송률이 2Mbits/s일 때의 통신 에너지 효율 $E_c = 2 \times 10^6$ [bits/J]을 사용하였다[8]. 실험 결과 신경망 분할 지점에 따라 종단 간 에너지 소모량이 유의미한 차이를 보이는 것을 확인할 수 있었으며, 제안 기법에 따라 5번째 블록에서 인공 신경망을 분할하는 것이 에너지 소모량을 최소화하는 방법임을 확인할 수 있다.

특히, 제안하는 신경망 분할 기반 협력 추론 기법을 활용하는 경우 모바일 단말에서 전체 추론 연산을 수행하는 것에 비해 약 2.5배, 에지 서버에서 전체 추론 연산을 수행하는 것에 비해 약 3배 이상 적은 에너지로 인공 신경망 추론 연산을 수행할 수 있다. 즉, 제안하는 신경망 분할 기반 협력 추론 기법을 활용하는 경우 동일한 용량의 배터리를 탑재한 모바일 단말에서 최대 3배 많은 인공 신경망 추론을 수행할 수 있음을 알 수 있으며, 결과적으로 제안하는 신경망 분할 기반 협력 추론 기법이 모바일 단말에서의 인공 신경망 추론의 에너지 효율을 대폭 개선할 수 있음을 확인할 수 있다.

V. 결론 및 향후 연구

본 논문에서는 신경망 분할 및 에지 서버와의 협력을 통해 모바일 단말에서 인공 신경망 추론의 에너지 효율을 개선하는 신경망 분할 기반 협력 추론 기법을 제안하였다. 또한, 협력 추론 기법의 에너지 소모량을 분석하고 이를 최소화할 수 있는 최적의 신경망 분할 지점을 제시하였다. 뿐만 아니라, 모의 실험을 통해 제안하는 기법을 현실적인 시스템 환경에서 실제 인공 신경망의 추론에 사용했을 때 에너지 효율을 개선할 수 있음을 확인하였다. 향후 연구 방향으로, 실측

을 통해 유휴 상태 및 데이터 로딩에 소요되는 전력 소모량을 분석하여 추론 지연시간을 고려한 인공 신경망 추론의 에너지 효율 최적화에 관한 연구를 수행할 예정이다.

References

[1] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270-64277, Oct. 2018. (https://doi.org/10.1109/ACCESS.2018.2877890)

[2] S. Yun, J.-M. Kang, S. Choi, and I.-M. Kim, "Cooperative inference of DNNs over noisy wireless channels," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8298-8303, Aug. 2021. (https://doi.org/10.1109/TVT.2021.3092179)

[3] S. Yun, W. Choi, and I.-M. Kim, "Cooperative inference of DNNs for delay-and memory-constrained wireless IoT systems," *IEEE Internet of Things J.*, vol. 9, no. 17, pp. 16113-16127, Sep. 2022. (https://doi.org/10.1109/JIOT.2022.3152359)

[4] B.-J. Choi, S. Yun, S.-H. Lee, and J.-M. Kang, "Analysis on inference latency of recurrent neural networks with split computing," *J. KIIS*, accepted, 2023. (https://doi.org/10.1109/ACCESS.2018.2877890)

[5] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," *ACM SIGARCH Comput. Archit. News*, vol. 45, no. 1, pp. 615-629, Apr. 2017. (https://doi.org/10.1145/3037697.3037698)

[6] H. Liu, L. Cao, T. Pei, Q. Deng, and J. Zhu, "A fast algorithm for energy-saving offloading with reliability and latency requirements in multi-access edge computing," *IEEE Access*, vol. 8, pp. 151-161, Dec. 2019. (https://doi.org/10.1109/ACCESS.2019.2961453)

[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. CVPR 2018*, pp. 4510-4520, Salt Lake City, USA, Jun. 2018. (https://doi.org/10.1109/CVPR.2018.00474)

[8] F. Kaup, P. Gottschling, and D. Hausheer, "PowerPi: Measuring and modeling the power consumption of the Raspberry Pi," in *Proc. 39th Annu. IEEE Conf. LCN 2014*, pp. 236-243, Edmonton, Canada, Sep. 2014. (https://doi.org/10.1109/LCN.2014.6925777)