인공 신경망 협력 추론 가속화 기법에 관한 연구

윤상석
부경대학교
ssyun@pknu.ac.kr

A Study on Accelerating Cooperative Inference of Artificial Neural Networks

Sangseok Yun
Pukyong National University

요 약

본 논문에서는 인공 신경망의 협력 추론 가속화를 위한 기법에 관해 연구하였다. 특히, 협력 단말 간 무선 채널을 오류가 존재하지 않는 완벽 채널로 가정하는 대부분의 기존 협력 추론 관련 연구와 달리 본 연구에서는 오류가 존재하는 현실적인 무선 채널을 고려하였으며, 이러한 환경에서 협력 추론에 소요되는 지연 시간을 최소화할 수 있는 협력 추론 가속화 기법을 연구하였다.

Ⅰ. 서 론

최근 인공지능 기술의 발전에 힘입어 다양한 IoT (Internet of Things) 기기에서 심층 인공 신경망을 사용하고자 하는 연구가 활발히 수행되고 있다. 하지만 IoT 단말들은 일반적으로 연산 능력이 제한되어 있고, 따라서 심층 인공 신경망의 추론 연산 수행이 과도한 지연 시간을 유발하는 등의 문제를 유발할 수 있다. 이러한 문제를 해결하기 위해 최근 IoT 단말과 고성능 서버가 심층 인공 신경망의 추론을 위해 협력하는 협력 추론 기법이 활발히 연구되고 있다[1]. 하지만 대부분의 기존 협력 추론 기법 관련 연구는 IoT 단말과 고성능 서버 사이의 채널을 오류가 존재하지 않는 완벽 채널로 가정하고 있어 실용적이지 않다는 문제를 가진다[2].

본 연구에서는 기존 협력 추론 관련 연구의 이러한 한계를 극복하기 위해 IoT 단말과 고성능 서버 간의 무선 채널을 오류가 존재하는 현실적인 무선 채널로 모델링 하였으며, 이러한 환경하에서 협력 추론에 소요되는 지연 시간을 최소화할 수 있는 협력 추론 기법을 개발하였다.

Ⅱ. 협력 추론

협력 추론은 IoT 단말에서 수행되는 심층 신경망의 추론이 과도한 연산 지연 시간을 유발하는 문제를 극복하기 위해 사용되는 기술이다. 협력 추론 기법을 사용하는 경우, 그림 1과 같이 심층 인공 신경망 추론 과정의 일부는 IoT 단말에서 수행하고 추론 과정의 나머지는 높은 연산 능력을 가진 서버에서 수행함으로써 심층 인공 신경망 추론의 지연 시간을 저감할 수 있다. 다시 말해서, 1) IoT 단말에서는 심층 인공 신경망의 ℓ번째 레이어까지의 추론을 수행한 후 2) 중간 결과, 즉 ℓ번째 레이어의 출력을 무선 채널을 통해 고성능 서버로 전송한다. 이어서 3) 높은 연산 능력을 가진 서버는 에지 단말로부터 수신한 중간 결과를 심층 인공 신경망의 $\ell+1$번째 레이어의 입력으로 사용하여 $\ell+1$번째 레이어부터의 추론 과정을 빠른 속도로 수행하고 4) 최종 추론 결과를 IoT 단말로 전달한다. 따라서, 무선 협력 추론을 위한 종단 간 (end-to-end) 지연 시간은 전술한 1)과 3)에서 소요되는 연산 지연 시간과 2)와 4)에서 소요되는 통신 지연 시간의 합으로 나타낼 수 있다.

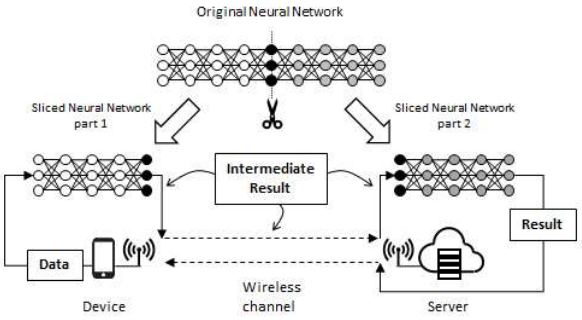

그림 1 무선 협력 추론 기법의 개념도

IoT 단말과 고성능 서버에서 심층 인공 신경망의 i번째 계층의 연산을 수행하는 데 소요되는 지연 시간을 각각 $T_d(i)$ 그리고 $T_s(i)$라 하고 추론하고자 하는 심층 인공 신경망을 구성하는 계층의 수를 L이라 하면, 1)과 3)에서 소요되는 연산 지연 시간의 합 $D_{comp}(\ell)$은 다음 수식 (1)과 같이 나타낼 수 있다.

$$D_{comp}(\ell) = \sum_{i=1}^{\ell} T_d(i) + \sum_{i=\ell+1}^{L} T_s(i) \quad (1)$$

한편, 무선 협력 추론의 2)에서는 무선통신을 활용해 ℓ번째 계층의 출력 feature vector를 고성능 서버로 전송해야 하며, 4)에서는 무선통신을 활용해 최종 출력을 IoT 단말에게 피드백해야한다. 그런데, 일반적으로 인공 신경망의 최종 출력은 은닉 계층의 출력 feature vector에 비해 무시할만큼 작은 데이터의 크기를 가진다. 예를 들어, 인공 신경망이 가장 성공적으로 적용되는 이미지 분류 문제에서 은닉 계층의 출력 feature vector는 이미지의 크기에 비례하는 데이터 크기를 가지지만, 최종 출력은 입력 이미지의 레이블로 수 내지 수십 비트 수준의 작은 데이터 크기를 가진다[3]. 따라서, 4)에 소요되는 통신 지연을 무시하고 2)에 소요되는 통신 지연만을 고려하면, 협력 추론의 통신 지연 $D_{comm}(\ell)$은 다음 수식 (2)와 같이 나타낼 수 있다.

$$D_{comm}(\ell) = B(\ell)/RW \quad (2)$$

여기서 $B(\ell)$, R, W는 각각 ℓ번째 계층의 출력 feature vector를 나타내는 데 필요한 비트의 수, 무선통신의 전송률 및 대역폭을 나타낸다.

본 연구에서는 IoT 단말과 고성능 서버 간의 무선 채널을 잡음이 존재하는 현실적인 채널로 모델링한다. IoT 단말과 고성능 서버 간의 채널 이득과 대역폭을 h라 하고 송신 전력과 수신 단의 잡음의 전력의 비, 즉 SNR (signal-to-noise ratio)을 γ라고 하면, IoT 단말과 고성능 서버 간의 채널 용량 C를 아래 수식 (3)과 같이 나타낼 수 있다.

$$C = \log_2(1+|h|^2\gamma) \tag{3}$$

정보이론에 따르면 무선통신의 전송률 R이 채널 용량 C보다 작은 경우에만 오류 없는 정보 전송이 가능하며, 따라서 IoT 단말과 고성능 서버 사이의 데이터 전송은 정전 확률 $P_o(R) = \Pr(C < R)$만큼의 확률로 실패할 수 있어 재전송이 필요할 수 있다. 무선 채널 이득 h가 평균이 0이고 분산이 σ_h^2인 가우시안 분포를 따른다고 하면, 전송률 R에 따른 정전 확률 $P_o(R)$을 아래 수식 (4)와 같이 나타낼 수 있다.

$$P_o(R) = 1 - \exp\left(-\frac{2^R-1}{\sigma_h^2\gamma}\right) \tag{4}$$

따라서, 주어진 채널 환경하에서 평균적으로 획득할 수 있는 IoT 단말과 고성능 서버 간의 실효 전송률 (throughput)은 $\overline{R} = R(1-P_o(R))$이며, 협력 추론의 평균 통신 지연 $\overline{D}_{comm}(\ell,R) = B(\ell)/\overline{R}W$이고, 결과적으로 목표 전송률에 따른 정전 확률을 고려한 협력 추론의 평균 종단 간 지연 시간 $\overline{D}(\ell,R) = D_{comp}(\ell) + \overline{D}_{comm}(\ell,R)$이 된다.

Ⅲ. 협력 추론 가속화 기법

현실적인 무선 채널을 고려한 협력 추론의 평균 종단 간 지연 시간을 최소화하기 위해서는 신경망 분할 지점 ℓ과 무선통신의 목표 전송률 R을 아래 수식 (5)와 같이 공동 최적화해야 한다.

$$(\ell^*, R^*) = \min_{\substack{1 \le \ell \le L \\ R \ge 0}} \overline{D}(\ell,R) \tag{5}$$

이때, 무선통신의 목표 전송률 R은 협력 추론의 연산에 소요되는 지연 시간 $D_{comp}(\ell)$과는 무관하며 평균 통신 지연 시간 $\overline{D}_{comm}(\ell)$의 분모에만 영향을 미치는 값이다. 따라서 목표 전송률 R의 최적값은 신경망 분할 지점 ℓ과는 독립적으로 결정된다. 따라서 본 연구에서는 최적의 목표 전송률 R^*을 먼저 계산한 후 이를 바탕으로 최적의 신경망 분할 지점 ℓ^*을 결정하는 순차적 최적화 방법을 제안한다.

먼저 최적 목표 전송률 R^*은 협력 추론의 평균 통신 지연 $\overline{D}_{comm}(\ell)$을 최소화하는 목표 전송률이며, 결과적으로 IoT 단말과 고성능 서버 간의 실효 전송률 $\overline{R} = R(1-P_o(R))$을 최대화하는 값이다. 따라서 최적의 목표 전송률 R^*은 아래 수식 (6)과 같이 계산할 수 있다.

$$R^* = \max_{R \ge 0} R\exp\left(-\frac{2^R-1}{\sigma_h^2\gamma}\right) = \frac{\omega(\sigma_h^2\gamma)}{\ln(2)} \tag{6}$$

이때, $\omega(\cdot)$는 Lambert W 함수이며, 수식 (6)의 마지막 등호는 Lambert W 함수의 정의에 의해 성립한다. 또한, 주어진 R^* 값 하에서 최적의 신경망 분할 지점 $\ell^* = \min_{1 \le \ell \le L} \overline{D}(\ell,R^*)$이며, 결과적으로 제안하는 협력 추론 가속화 기법을 사용했을 때 획득할 수 있는 최적의 평균 종단 간 지연 시간은 $\overline{D}(\ell^*,R^*)$이다.

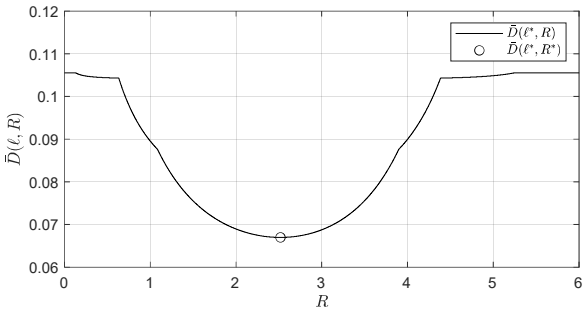

그림 2 목표 전송률에 따른 협력 추론의 평균 종단간 지연 시간

Ⅳ. 모의실험

본 연구에서 제안한 협력 추론 가속화 기법의 성능을 검증하기 위해 고성능 데스크탑 PC와 연산 능력을 의도적으로 저하시킨 구형 노트북을 이용해 고성능 서버와 IoT 단말 간의 협력 추론 모의실험을 수행하였다. 모의실험에 사용한 인공 신경망은 CIFAR-10 데이터셋을 위한 ResNet[4]이며, $\gamma = 10$ dB, $W = 10$ MHz, 그리고 $\sigma_h^2 = 1$을 활용하였다. 목표 전송률에 따른 협력 추론 기법의 평균 종단간 지연 시간을 나타낸 그림 2를 통해 목표 전송률 최적화가 평균 종단간 지연 시간에 큰 영향을 주는 것을 확인할 수 있으며, 제안한 협력 추론 가속화 기법 사용 시 최대 36.5% 이상의 지연 시간을 저감할 수 있음을 확인할 수 있다.

Ⅴ. 결론

본 연구에서는 현실적인 무선 채널 환경을 고려한 심층 인공 신경망의 무선 협력 추론에 관해 연구하였다. 특히, 주어진 무선 채널 환경하에서 최적의 무선통신 전송률을 계산함으로써 협력 추론을 사용할 때 발생하는 종단 간 지연 시간을 최소화할 수 있는 기법을 개발하였다. 모의실험 결과 제안한 협력 추론 가속화 기법을 사용하는 경우 기존 대비 평균 종단 간 지연 시간을 획기적으로 저감할 수 있음을 확인하였다.

ACKNOWLEDGMENT

이 논문은 4단계 BK21 사업(스마트로봇융합응용교육연구단)에 의하여 지원되었음.

참 고 문 헌

[1] 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.

[2] A. E. Eshratifar, M. S. Abrishami, and M. Pedram, "JointDNN: An efficient training and inference engine for intelligent mobile cloud computing services," *IEEE Trans. Mobile Comput.,* vol. 20, no. 2, pp. 565 - 576, Feb. 2021.

[3] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR 2009*, Miami, USA, Jun. 2009, pp. 248 - 255.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR 2016,* Las Vegas, USA, Jun. 2016, pp. 770 - 778.