Split Computing 연구 동향

김보경, 고한얼
경희대학교 전자정보융합공학과
{bokyeong0405, heko}@khu.ac.kr

Trends on the Split Computing

Bokyeong Kim and Haneul Ko
Department of Electronics and Information Convergence Engineering, Kyung Hee University

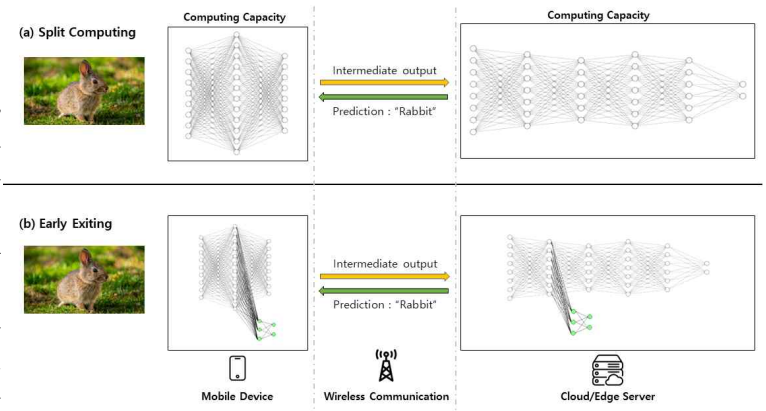
요 약

최근 딥러닝(DL) 분야는 컴퓨터 비전, 자연어 처리, 디지털 신호 처리, 무선 네트워킹과 같은 영역에서 놀라운 속도로 발전했다. 이에 따라 딥러닝을 스마트폰의 음성 인식, 실시간 무인 내비게이션 및 드론 기반 감시와 같은 모바일 응용 프로그램에 적용하려는 시도가 있었다. 그러나 모바일 장치의 제한된 컴퓨팅 자원 때문에 NN 추론 작업에 지나치게 긴 시간이 소요되어 모바일 장치에서 실행하기에 한계가 있다. 모바일 장치에 할당된 계산 부담을 덜어주기 위해 split computing(SC)이 고안되었고, NN 추론 작업에 필요한 총 소요 시간을 획기적으로 줄이기 위해 early exiting(EE) 전략이 고안되었다. 따라서 본 논문에서는 SC와 EE의 최신 연구 동향을 정리하여 향후 어떤 연구가 추가로 필요할지를 고찰한다.

Ⅰ. 서론

딥러닝(DL) 분야는 지난 몇 년 동안 컴퓨터 비전(CV), 자연어 처리(NLP), 디지털 신호 처리 (DSP), 무선 네트워킹[1, 2]과 같은 영역에서 놀라운 속도로 발전했다[3]. 오늘날의 최첨단 심층 신경망(DNN)은 수천 개의 이미지를 높은 정확도로 분류할 수 있고[4], 최첨단 심층 강화 학습(DRL)은 아타리 게임[5]이나 바둑[6]의 수많은 복잡한 최적화 작업에서 사람에 가까운 성능을 제공하는 것으로 나타났다. 딥러닝의 예측 정확도가 향상함에 따라 이를 스마트폰의 음성 인식[7, 8], 실시간 무인 내비게이션[9] 및 드론 기반 감시[10, 11]와 같은 모바일 응용 프로그램의 복잡한 추론 작업 수행에 적용하려는 연구가 진행되었다. 그러나 대부분의 모바일 장치는 복잡한 DNN 모델의 계산 요구사항을 만족하지 못한다. 이 문제를 해결하기 위해 모바일에서 얻은 input을 클라우드 서버로 보내 전체 NN 추론 작업을 클라우드 서버에서 진행하는 클라우드 기반 접근 방식이 제안되었다.

클라우드 기반 접근 방식은 모바일 장치에서 계산하지 않아도 되고, 모바일 장치에서 계산할 때보다 계산 속도가 빠르다. 하지만 전체 input 데이터를 보내기 때문에 보안성 문제가 있고, 데이터 전송 지연이 발생할 수 있다는 문제가 발생할 수 있다. 여기서 모바일 장치부터 클라우드까지의 데이터 전송 지연을 줄이기 위해 에지 컴퓨팅(EC) 기법을 도입할 수 있다. EC는 심층 신경망을 모바일 장치와 매우 가까운 서버, 즉 네트워크의 "에지"에 위치한 서버로 모바일 장치의 작업을 완전히 오프로드 하는 방법을 사용한다. 이를 통해 모바일 장치의 부족한 계산 자원 문제와 클라우드 기반 접근 방식에서의 전송 지연을 해결할 수 있다. 하지만 불규칙한 노이즈 및 간섭 패턴으로 인해 애플리케이션의 성능이 저하될 수 있다. 무선 채널의 불안정성과 일부 모바일 장치 계산 자원 부족 문제를 해결하기 위해서는 에지로 오프로드 되는 데이터의 양을 줄이는 동시에 정확도는 원래에 가깝게 유지해야 한다. 이러한 이유에서 input 데이터에 대한 보안성을 보장하기 위해 NN을 특정 layer를 기준으로 head와 tail로 쪼개어 input 데이터가 필요한 head는 모바일 장치에서, tail은 클라우드 서버에서 계산하는 split computing(SC) [12]이 제안되었다. 또한, 총 추론 시간을 획기적으로 줄이기 위해 NN의 특정한 hidden layer에 output을 생성할 수 있는 여러 개의 출구를 배치하여 목표 신뢰도를 달성하는 출구를 선택해 그 이후의 layer에 대한 계산은 생략하는 early exiting(EE)[13]이 제안되었다.



〔그림 1〕 (a)split computing, (b)early exiting 개요:이미지 분류 예제

본 논문의 나머지 구성은 다음과 같다. 2장 본론에서 split computing과 early exiting 모델을 소개하고 해당 기술을 활용한 연구를 요약한다. 3장 결론에서는 본 연구의 의의를 정리하고 향후 연구 방향을 제시하며 논문을 마친다.

Ⅱ. 본론

에지 컴퓨팅의 경우 전체 input 데이터를 에지 서버로 오프로드해 모바일 장치가 추론 작업을 수행하지 않아도 되며 정확도를 보존할 수 있다는 장점이 있다. 하지만 input 데이터가 매우 큰 경우 무선 네트워크를 통해 데이터를 전송하는데 긴 시간이 걸리게 되며 전송 에러의 위험이 있고, 에지가 전체 input 데이터를 알기 때문에 보안성 문제가 발생할 수 있다. 로컬 컴퓨팅의 경우 input 데이터를 에지로 보내지 않아도 된다는 장점이 있지만, 모바일 장치는 계산 능력이 떨어져 추론 작업을 마치는데 긴 시간이 필요하다. 따라서, input 데이터를 에지나 클라우드에 전송하지 않고 모바일 장치의 계산 부담을 줄여주기 위해 SC와 EE가 제안되었다.

그림 1은 이미지 분류 예제에서 SC와 EE의 개요를 보여준다. SC와 EE 모두 모바일 기기에 있는 이미지를 input으로 하고 DNN 추론 결과를

output으로 한다. SC는 NN 추론 과정을 모바일 장치와 클라우드/에지 서버에 분산시키고, 중간 데이터를 전송할 때 걸리는 시간을 줄이는 것을 목표로 한다. SC는 그림 1 (a)에서처럼 특정 layer를 기준으로 NN을 head와 tail로 나눠 각각 모바일 장치와 클라우드/에지 서버에서 추론하는 방식이다. 이 방식을 사용하면 모바일 장치에서 head에 대한 추론 작업을 수행하기 때문에 input 데이터를 에지/클라우드 서버로 보내지 않고 head 추론 작업의 output(중간 데이터)만 전송하면 된다. 이 방식은 추론 정확도는 유지할 수 있지만 NN 추론 작업의 일부를 클라우드/에지 서버보다 계산 능력이 떨어지는 모바일 장치에서 수행하기 때문에 총 추론 시간이 클라우드/에지 서버에서 전체 NN 추론 작업을 수행할 때보다 더 클 수 있다. SC의 총 추론 시간을 최소화하는 문제의 핵심은 중간 데이터 전송 시간을 줄이는 것이다. 최근에는 이 중간 데이터 전송 시간을 줄이기 위해 head에서 output layer의 node 개수를 줄이거나 양자화를 적용해 output 데이터 개수를 줄이는 연구가 진행되었다[14].

EE의 핵심 아이디어는 DNN에 출구(early exit)를 도입해 DNN 모델을 작게 만들어야 할 필요 없도록 하는 것이다. 일반적으로 추론 시간을 줄이기 위해 DNN 모델을 작게 만드는데, EE는 NN 모델 전체에 걸쳐 NN에 대한 output을 만들어낼 수 있는 출구를 여러 개 배치한 모델을 만들어 목표로 한 신뢰도를 만족하는 첫 번째 출구를 선택하는 방식을 통해 추론 시간을 단축한다. 예를 들어 그림 1 (b)에서 에지/클라우드 서버에 위치한 출구가 output에 대한 충분한 신뢰도를 가진다고 가정했을 때, 해당 지점에서 추론 작업을 종료해 다음 layer가 실행되지 않기 때문에 추론 작업 시간을 단축할 수 있다. 하지만 원하는 신뢰도에 도달할 때까지 모든 출구에 대한 계산이 수행되기 때문에 만약 input layer부터 출구까지의 layer 수가 출구 이후부터 output layer의 layer 수보다 많으면 계산 복잡도가 증가해 전체 추론 비용이 오히려 평균적으로 증가하게 된다. 따라서 출구까지의 layer 수가 출구 이후의 layer 수보다 적어야 한다.

최근 모바일 장치와 클라우드 서버 사이에 여러 개의 에지를 두어 NN을 분산시켜 추론 작업을 수행하는 다중 분할 SC 연구[15], 모바일-에지-클라우드 컴퓨팅 시스템에 EE를 적용한 연구[16]가 진행되었다.

Ⅲ. 결론 및 향후 연구

본 논문에서는 계산 자원이 한정적인 모바일 장치에 대한 DNN 추론 작업을 수행할 수 있도록 하는 기술인 split computing과 early exiting에 대해 요약하고 선행 연구를 정리했다. 향후 split computing의 중간 데이터 전송 지연을 줄여 전체 NN 추론 작업 시간을 줄이는 연구를 진행할 예정이다.

참 고 문 헌

[1] J. Jagannath, N. Polosky, A. Jagannath, F. Restuccia, and T. Melodia, "Machine learning for wireless communications in the Internet of Things: A comprehensive survey," *Ad Hoc Networks*, vol. 93, no. 10, pp. 1-46, Jun. 2019.

[2] F. Restuccia and T. Melodia, "Deep Learning at the Physical Layer: System Challenges and Applications to 5G and Beyond," *IEEE Communications Magazine*, vol. 58, no. 10, pp. 58-64, Oct. 2020.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning.," *Nature*, vol. 521, no. 7553, pp. 436-444, May. 2015.

[4] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, ... and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1-36, Sep. 2018.

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," in *Proc. NIPS Deep Learning Workshop 2013*, Dec. 2013.

[6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, ... and D. Hassabis, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354-359, Oct. 2017.

[7] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP) 2013*, Oct. 2013.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, ... and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, Oct. 2012.

[9] R. P. Padhy, S. Verma, S. Ahmad, S. K. Choudhury, and P. K. Sa, "Deep Neural Network for Autonomous UAV Navigation in Indoor Corridor Environments," *Procedia Computer Science*, vol. 133, pp. 643-650, Jul. 2018.

[10] A. Singh, D. Patil, and S. N. Omkar, "Eye in the Sky: Real-Time Drone Surveillance System (DSS) for Violent Individuals Identification Using ScatterNet Hybrid Deep Learning Network," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018*, Jun. 2018.

[11] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang, "Person Re-Identification in Aerial Imagery," *IEEE Transactions on Multimedia*, vol. 23, pp. 281-291, Mar. 2020.

[12] 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 Computer Architecture News*, vol. 45, no. 1, pp. 615-629, Mar. 2017.

[13] S. Teerapittayanon, B. McDanel, and H. T. Kung, "BranchyNet: Fast inference via early exiting from deep neural networks," in *Proc. International Conference on Pattern Recognition (ICPR) 2016*, Dec. 2016.

[14] Y. Matsubara, S. Baidya, D. Callegaro, M. Levorato, and S. Singh, "Distilled Split Deep Neural Networks for Edge-Assisted Real-Time Systems," in *Proc. International Conference on Mobile Computing and Networking(MobiCom) 2019*, Oct. 2019

[15] S. Wang, X. Zhang, H. Uchiyama, and H. Matsuda, "HiveMind: Towards Cellular Native Machine Learning Model Splitting," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 2, pp. 626-640, Oct. 2021.

[16] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices," in *Proc. IEEE 37th International Conference on Distributed Computing Systems (ICDCS) 2017*, Jul. 2017.