다중 사용자 모바일 비전 어플리케이션을 위한 동적 딥러닝 모델 및 자원 스케일링 방안

최평준, 서정민, 곽정호
대구경북과학기술원
pyeongjun.choi@dgist.ac.kr, jeongminseo@dgist.ac.kr, jeongho.kwak@dgist.ac.kr

Dynamic Deep Learning Model and Resource Scaling Scheme for Multi-user Mobile Vision Application

Choi Pyeong Jun, Seo Jeong Min, Kwak Jeong Ho
DGIST

요 약

자율주행차나 증강현실과 같이 딥러닝을 활용하는 모바일 비전 어플리케이션에 대한 소비자의 수요가 증가하면서 학계와 산업계의 관심도 같이 커지고 있다. 딥러닝 기반 모바일 비전 어플리케이션은 사용하는 딥러닝 모델의 너비, 깊이 등 모델 측면의 특성과 사용하는 모바일 단말의 컴퓨팅 성능, 네트워크 환경, 발열 등 자원 측면의 특성 및 처지에 따라 그 성능이 크게 변한다. 기존 연구들은 대부분 주어진 단말에 대해서 딥러닝 모델을 최적화하거나 특정 모델에 대해서 동적 코드 오프로딩이나 동적 GPU 주파수 조절을 하는 등 자원을 최적화하는 단방향 최적화를 해왔다. 또한 모바일 단말의 네트워크 환경이나 컴퓨팅 자원, 유저의 서비스 요청량 등이 특정한 분포를 따르는 등 비현실적인 가정을 포함해 실제 환경과 동떨어진 부분이 있었다. 우리는 현실적인 상황에서도 적응적으로 딥러닝 모델과 단말, 엣지 서버의 자원을 동시에 최적화하여 단말의 프레임 처리량, 평균 정확도, 에너지 소모량, 발열, 메모리 용량 등 다양한 요구 조건을 만족할 수 있는 방안을 제안한다.

Ⅰ. 서 론

모바일 단말의 성능이 증가하고 가벼운 딥러닝 모델이 등장하면서 간단한 딥러닝 어플리케이션은 모바일 단말 내에서도 실행할 수 있게 됨에 따라 딥러닝을 활용한 다양한 모바일 비전 어플리케이션이 보편화되고 있다. 이런 모바일 서비스들의 성능은 동적으로 변하는 서비스 요청, 추론 단말의 컴퓨팅 자원 및 네트워크 환경과 추론에 사용하는 딥러닝 모델에 따라 달라질 수 있다. 예를 들어, 풍부한 컴퓨팅 자원과 대규모 딥러닝 모델을 가진 엣지 서버로 오프로딩을 한다면 높은 정확도 성능과 빠른 추론 속도를 얻을 수 있지만 네트워크 속도에 따라 추가적인 지연이 발생할 것이고, 반대로 소규모 딥러닝 모델과 부족한 자원을 가진 모바일 단말 내부에서 추론을 한다면 낮은 정확도와 느린 추론 속도를 얻을 것이다. 이런 서비스의 또 다른 문제점으로는, 모바일 단말 내부에서 추론을 하거나 엣지 서버로 대용량의 데이터를 전송할 때 발생하는 열과 배터리 소모가 있다. 특히 발열과 배터리 소모는 요청하는 서비스의 특성 (지연, 정확도, FPS), 단말의 특성 (프로세서 성능, 배터리 용량, 메모리 용량), 주변 환경에 따라 크게 바뀌므로 예측하기 어렵다. 서비스의 정확도, 지연, 단말의 발열과 배터리 소모 등은 사용자 경험에 직접적인 영향을 미치기 때문에 서비스 제공자는 쾌적한 서비스 제공을 위해 단말의 에너지 소모와 발열 관리, 서버의 부하 관리 등에 많은 비용을 사용하게 된다. 이에 정해진 자원 내에서 안정적인 서비스 제공을 위해 주어진 딥러닝 모델에 대해서 변화하는 모바일 단말의 환경에 적응적으로 단말과 엣지 서버의 컴퓨팅 자원 및 네트워킹 자원을 동적으로 최적화하는 연구들이 있었다 [1].

한편, 주어진 단말에 적합하도록 딥러닝 모델을 개선하는 연구도 있었다. 이들은 단말의 자원에 따라 필터의 수나 입력층의 크기 등 딥러닝 모델의 구조를 수정하거나 완성된 딥러닝 모델의 노드나 가중치 일부를 제거하는 방식으로 연산양과 정확도 성능을 조절했다 [2]. 하지만 단말에 맞게 딥러닝 모델을 최적화하는 연구들은 단말의 컴퓨팅 자원이 일정할 것이라는 가정을 필요로 한다. 딥러닝 모델은 학습이 완료되는 순간에 연산양과 정확도 성능 등의 특징이 고정되기 때문이다. 때문에 변화하는 단말의 자원

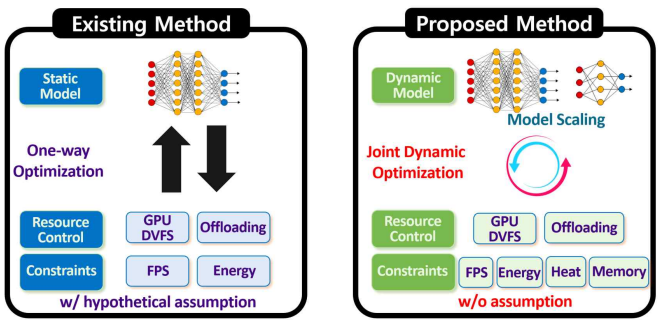

그림 1. 기존의 단방향 최적화 방안과 제안하는 동적 최적화 방안

상황에 동적으로 대응하지 못해 사용자에게 최적의 성능을 제공하지 못하게 된다. 이에 실시간으로 모델의 노드를 줄이거나, 하나의 딥러닝 모델에 여러 서브 모델을 탑재해서 가용 자원양에 따라 동적으로 모델을 선택할 수 있도록 하는 연구들이 있었다. [3, 4]

정리하자면 기존에는 그림. 1의 왼쪽과 같이 이미 완성된 모델에 대해 어떻게 자원을 활용할지를 결정하거나, 주어진 자원에 대해 딥러닝 모델을 최적으로 디자인하는 단방향 최적화 연구들이 진행되었다. 우리는 동적으로 변화하는 환경과 사용자의 요구량에 맞게 딥러닝 모델과 단말, 엣지 서버의 의 컴퓨팅/네트워킹 자원을 동시에 동적으로 최적화하는 아이디어를 고안했다. 제안하는 방안은 매 타임슬롯마다 단말의 에너지 소모, 발열, 메모리 사용량 등을 고려하여 최적의 딥러닝 모델과 단말의 GPU 클럭 주파수, 오프로딩 정책, 네트워크 속도, 서버의 자원 할당을 동적으로 제어한다.

Ⅱ. 시스템 모델, 알고리즘

우리는 그림 2와 같이 엣지 서버와 무선으로 연결되어 있는 여러 모바일 단말이 비전 어플리케이션을 수행하는 상황을 고려한다. 매 타임슬롯마다 비전 워크로드가 모바일 단말로 들어올 때, 단말은 비전 어플리케이션의 목표 정확도와 프레임 처리량, 에너지 소모량, 발열, 메모리 용량을 고려하여 오프로딩 여부와 단말의 GPU 주파수, 딥러닝 모델, 엣지 서버로의 전

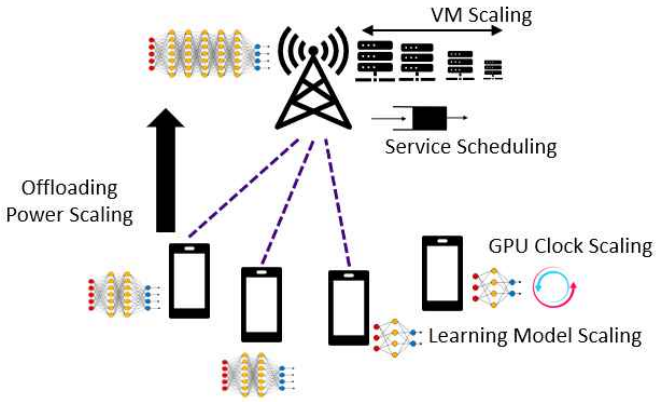

그림 2. 다중 유저를 위한 딥러닝 모델, 자원의 동적 스케일링 방안

송량을 결정한다. 엣지 서버는 사용자의 서비스 요구량과 서버 내 서비스 대기열을 고려하여 사용할 자원량과 유저 스케줄링을 진행한다. 우리의 목적은 위의 결정 변수들을 동적, 동시에 제어하여 평균 정확도를 목표 정확도 수준으로 유지하면서 프레임 처리량, 단말의 에너지 소모량, 발열, 메모리 용량을 모두 고려한 비용 함수를 최소화하는 것이다.

Ⅲ. 결론

본 논문에서는 딥러닝 모델과 컴퓨팅/네트워킹 자원을 동시에 동적으로 제어하여 목표 정확도를 만족하면서 프레임 처리량, 단말의 에너지 소모량, 발열, 메모리 용량을 최적화하는 방안을 제안했다. 단말의 환경에 따라 다양한 결정 변수를 동적으로 최적화하여 전천후 제어를 할 수 있을 것으로 기대되어 점점 더 복잡해져가는 인공지능 서비스 산업에서 중추적인 역할을 할 것으로 기대된다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1C1C1003030).

참 고 문 헌

[1] Galanopoulos, A., Ayala-Romero, J. A., Leith, D. J., & Iosifidis, G. (2021, May). AutoML for video analytics with edge computing. In IEEE INFOCOM 2021-IEEE Conference on Computer Communications (pp. 1-10). IEEE.

[2] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[3] Lin, J., Rao, Y., Lu, J., & Zhou, J. (2017). Runtime neural pruning. Advances in neural information processing systems, 30.

[4] Lou, W., Xun, L., Sabet, A., Bi, J., Hare, J., & Merrett, G. V. (2021). Dynamic-ofa: Runtime DNN architecture switching for performance scaling on heterogeneous embedded platforms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3110-3118).