엣지 컴퓨팅 기반 모바일 비전 AI 를 위한 DNN 모델 분할 및 해상도 최적화

박준수, 김영진*
인하대학교, *인하대학교

azfk008@gmail.com, *yj.kim@inha.ac.kr

요 약

딥러닝 네트워크(DNN) 모델을 기반으로 한 비전 애플리케이션의 사용자 경험 품질(QoE)은 모바일 기기의 하드웨어 성능, 변화하는 네트워크 상태, 그리고 DNN 모델의 특성에 의해 영향을 받는다. 본 논문에서는 시스템의 동적 변화에 대응하기 위해 DNN 모델 분할과 프레임 해상도를 확률적 최적화 기법을 통해 동적으로 최적화하는 *Parecon* 알고리즘을 소개한다. *Parecon* 은 *(i)* 모바일 기기와 MEC 서버 간의 분할 지점, *(ii)* 입력 프레임 크기, *(iii)* 각 시간 슬롯에서 처리할 프레임 수를 결정한다. 우리는 *Parecon* 이 이전 연구에서 공동으로 다루어지지 않았던 주요 QoE 지표인 단대단 지연 시간, 정확도, 처리량을 동시에 최적화한다는 것을 이론적으로 입증한다. 더불어, Nvidia Jetson Xavier NX 와 Nvidia RTX 4090 GPU 가 장착된 MEC 서버를 사용한 시뮬레이션과 테스트베드를 통해 기존 알고리즘과 비교한 *Parecon* 의 효과성과 우수성을 검증한다.

Ⅰ. 서 론

최근 통신 및 딥러닝 기술의 발전에 따라 자율주행 차량, 드론, 스마트폰과 같은 기기에서의 모바일 비전 애플리케이션 사용이 증가하고 있다. DNN 모델을 모바일 기기에서만 실행하면 모바일 성능 제한으로 인해 실행 속도가 느려질 수 있으며, DNN 모델의 실행을 전적으로 엣지 서버에 맡기게 되면, 네트워크의 상태가 좋지 않을 경우 통신 지연으로 인해 실행 속도가 늦어질 수 있다.

이 문제를 해결하기 위해 모바일 엣지 컴퓨팅(MEC)이라는 접근 방식이 필요하다. DNN 모델은 여러 계층으로 구성되어 있기 때문에, 이 접근 방식을 사용하면 모델을 여러 개의 계층 그룹으로 분할할 수 있다. 상황에 따라 일부 계층 그룹은 모바일 기기에서 처리되고, 중간 결과는 MEC 서버로 전송되어 나머지 계층을 처리한다. DNN 모델 분할에 대한 연구는 2017 년에 시작되었으며 [1], 이후 다양한 방법이 탐구되어 왔다. 그러나 DNN 모델 분할 과정에서 입력 이미지의 크기 조절을 통해 정확도를 타협해가며 타 성능을 향상시키는 연구가 아직 이루어지지 않았다.

DNN 모델 분할과 입력 프레임 크기 조정이 QoE (fps, 정확도, 단대단 지연 시간)에 끼치는 영향을 평가하기 위해, 우리는 모바일 기기 (Nvidia Jetson Xavier NX)와 Nvidia RTX 4090 GPU 가 장착된 엣지 서버를 이용해 사전 측정을 수행했다. 모바일 기기에서 이미지 분류에 사용되는 EfficientNetV2-S 모델을 두 기기에 배포하였고, EfficientNetV2-S 모델의 계층들을 9 개의 계층 그룹으로 분리했다. ImageNet 검증 세트에서 크기가 114x114 픽셀에서 384x384 픽셀까지 미리 조정된 1000 장의 이미지를 사용했다.

그림 1 은 입력 프레임 크기가 줄어들고 네트워크 상태가 좋아짐에 따라 fps 가 증가하고 단대단 지연 시간이 감소하는 것을 보여준다. 또한, 동일한 입력 프레임 크기와 네트워크 조건에서도 어떤 분할 지점을 선택하느냐에 따라 이러한 QoE 지표는 달라진다. 더불어, 그림 2(왼쪽)는 입력 프레임 크기가 줄어들수록 분류 정확도가 떨어진다는 것을 나타내며 이는 프레임 내 정보 손실로 인한 것이다. 그림 2(오른쪽)는 연속된 프레임 간의 컴퓨팅-네트워크 파이프라인이 약간의 단대단 지연 시간의 증가를 감수함으로써 fps 를 크게 향상시킬 수 있음을 보여주며, 그 효과는 선택된 분할 지점에 의해서도 영향을 받는다.

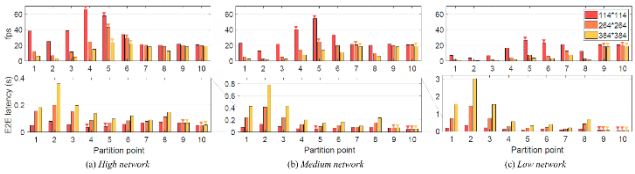

그림 1: 네트워크 상태와 분할 지점, 프레임의 크기가 fps 및 단대단 지연 시간에 미치는 영향.

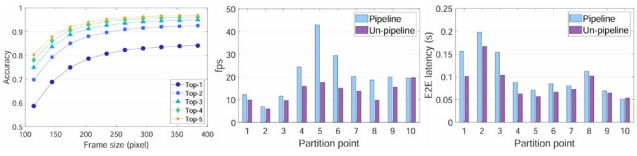

그림 2: (왼쪽) 입력 프레임의 크기가 정확도에 미치는 영향, (오른쪽) 파이프라인이 fps 및 단대단 지연 시간에 미치는 영향.

측정 결과를 바탕으로, 우리는 초 단위로 DNN 모델 분할 지점과 프레임 크기를 조정하는 *Parecon* 알고리즘을 제안한다. 이 알고리즘은 *(i)* 분할 지점, *(ii)* 처리 fps, *(iii)* 크기 조정 팩터를 동적으로 공동 결정한다.

이 논문의 기여는 다음과 같다: *1)* 모바일 비전 애플리케이션에서 분할 지점과 입력 프레임 크기 조정이 QoE 지표에 미치는 영향을 분석한다. *2)* 평균 단대단 지연 시간과 정확도에 대한 요구사항을 충족하면서 처리 fps 를 최대화하는 적응형 분할 및 해상도 제어 기법인 *Parecon* 을 제안한다. *3)* 실제 테스트베드를 기반으로 한 시뮬레이션과 실험을 통해 *Parecon* 이 기존 알고리즘보다 효과적이고 우수함을 입증한다.

Ⅱ. 본론

시스템 모델. 그림 3 은 우리가 설계한 시스템 구조를 보여준다. 모바일 기기와 MEC 서버는 사전 학습된 DNN 모델 m 을 실행하며, 모델 m 의 계층들은 N_m 개의 연속적인 계층 그룹으로 나뉜다. 계층 그룹 처리의 분할 지점은 $i(t) \in \{1,2,\cdots,N_m+1\}$ 로 정의된다. 이 분할 지점에 따라 *1)* 초기 계층 그룹은 모바일 기기에서 처리되고, *2)* 생성된 중간 결과는 무선 네트워크를 통해 MEC 서버로 전송되며, *3)* MEC 서버는 분할 지점 이후의 계층 그룹을 처리한 후, *4)* 최종 결과를 모바일 기기로

반환한다. 이 과정은 이 과정은 처리 fps 를 높이기 위해 연속적인 프레임 간의 파이프라인 방식으로 이루어진다.

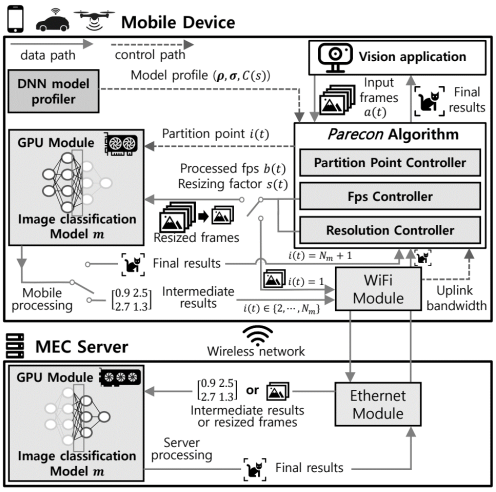

그림 3. MEC 서버의 지원을 받는 모바일 AI 시스템 구조.

각 시간 슬롯 t 에서, 모바일 기기는 비전 애플리케이션으로부터 $a(t)$ 개의 입력 프레임을 받는다. 모바일 기기는 처리할 프레임 수 $b(t) \in \{0, \cdots, a(t)\}$ 를 선택하고, 비율 $k(t) = b(t)/a(t)$ 를 계산한다. 각 프레임의 초기 크기는 w이며, DNN 모델 m에 삽입되기 전에 크기 조정 팩터 $s(t) \in \{s_1, s_2, \cdots, 1\}$에 의해 크기가 축소될 수 있다. 단대단 지연 시간 $R(t)$ 는 *(i)* 모바일 처리 시간, *(ii)* 중간 결과 업로드 시간, *(iii)* MEC 서버의 처리 시간 *(iv)* 최종 결과 다운로드 시간의 합으로 구성된다. 재조정된 프레임의 크기에 따라 해당 시간 슬롯 t동안 top-k 정확도 $C(s(t))$가 결정된다.

이 모델을 기반으로, 모바일 비전 애플리케이션의 세 가지 주요 QoE 지표인 처리 fps, 단대단 지연 시간, 정확도를 고려한 최적화 문제를 다음과 같이 수식화한다.

$$(\mathbf{P}): \max_{i,b,s} U(\bar{k}),$$

s.t. (C1): $\bar{R} \le r_{\text{th}}$,
(C2): $\overline{C(s)} \ge c_{\text{th}}$,
(C3): 네트워크 및 처리 자원 용량 제한,
(C4): 실행 가능한 도메인 내에서 $(i(t), b(t), s(t))$를 결정한다.

여기서 $U(\cdot)$는 모바일 사용자의 만족도를 나타내는 fps 유틸리티 함수이며, r_{th}는 목표 단대단 지연 시간, c_{th}는 목표 top-k 정확도이다.

알고리즘 제안. 우리는 가상 큐 기반 Lyapunov 최적화 기법과 convex 최적화를 기반으로 매 시간 슬롯마다 최적의 $(i(t)^*, b(t)^*, s(t)^*)$ 를 찾는 *Parecon* 알고리즘을 제안한다. 또한, *parecon* 의 복잡도를 분석하고 이론적인 성능 한계를 증명한다.

시뮬레이션 결과. 우리는 *Parecon* 의 성능을 평가하기 위해 트레이드 기반 시뮬레이션과 실제 실험을 수행하며, 최신 알고리즘들과 비교한다. 대전에서 측정한 LTE 와 5G 네트워크 트레이스를 사용했다. 모바일 기기, MEC 서버, DNN 모델, 이미지 데이터셋은 예비 측정에서 사용한 것과 동일하다. 시뮬레이터와 테스트베드는 Matlab 9.10 과 Python 3.8 을 기반으로 구현되었으며, 파이프라인 시스템은 Python 의 multiprocessing 모듈을 기반으로 구현되었다.

그림 4 는 목표 정확도 및 단대단 지연 시간과 일부 파라미터를 변경하면서 *Parecon* 의 시간 평균 처리 fps, 단대단 지연 시간, top-1 정확도에 대한 시뮬레이션 결과를 보여준다. 이러한 결과는 *Parecon*이 다양한 목표 요구 사항을 가진 여러 애플리케이션에 적용 가능함을 입증한다.

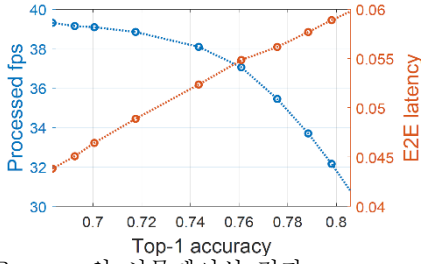

그림 4: *Parecon* 의 시뮬레이션 결과.

그림 5 는 *Parecon* 을 최신 알고리즘들과 비교한 실험 결과를 보여준다. 비교 대상은 *BinaryMEC* (해상도 제어를 포함한 이진 기반 코드 오프로딩), *SingleOpt* (단일 프레임에 최적화된 분할 및 해상도 제어), *CutEdge* (파이프라인을 고려한 해상도 제어 없이 최적화된 분할)이다. 우리는 목표 단대단 지연시간과 top-1 정확도를 각각 $r_{\text{th}} = 60\text{ms}$, $c_{\text{th}} = 0.79$ 로 설정한다. *Parecon* 은 목표 단대단 지연 시간과 정확도를 모두 만족하면서 가장 높은 처리 fps 를 달성한다.

그림 5: *Parecon* 및 최신 알고리즘들의 실험 결과.

Ⅲ. 결론

우리는 MEC 환경에서 모바일 비전 애플리케이션의 사용자 경험 품질을 향상시키기 위해 DNN 모델 분할과 입력 프레임 해상도를 공동으로 최적화하는 알고리즘을 개발했다. 심층적인 시뮬레이션을 통해 *Parecon* 알고리즘이 처리 fps, 단대단 지연 시간, 정확도 간의 균형을 최적으로 찾아냄을 입증했다.

ACKNOWLEDGMENT

본 연구는 2024 년도 정보통신기획평가원 (No.RS-2022-00155915 인공지능융합혁신인재양성(인하대학교), 2021-0-02201 사용자 프라이버시를 보존하는 비디오 캐싱을 위한 연합 학습 시스템, 2022-0-00448 인간처럼 회상이 가능한 인공 신경망 지속학습 플랫폼 개발, No. RS-2024-00398157 AI-Native 응용서비스 지원 AI 오케스트레이터 개발) 및 한국연구재단 (No. RS-2023-00240019)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] Y. Kang *et al.*, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," Proc. of ACM ASPLOS, p. 615-629, (2017).