스마트 카메라 엣지 클러스터에서 DNN 하이브리드 스케줄링 알고리즘

이찬민[1], 서민석[1], 박주성[1], 진민규[2], 박형빈[1], 이수경[1]
[1]연세대학교 컴퓨터과학과, [2]연세대학교 전기전자공학과

noel98@yonsei.ac.kr, tat1218@naver.com, pjs7153@yonsei.ac.kr,
wlsalsrb1236@yonsei.ac.kr, hbp@yonsei.ac.kr, sklee@yonsei.ac.kr

DNN Hybrid Scheduling Algorithm in Smart Camera Edge Cluster

Chan-Min Lee[1], Min-Seok Seo[1], Ju-Seong Park[1], Min-Gyu Jin[2], Hyung-Bin Park[1],Su-Kyoung Lee[1]
[1] Dept. of Computer Science, Yonsei University
[2]Dept. of Electrical & Electronic Engineering, Yonsei University

요 약

본 논문에서는 엣지 컴퓨팅에서 다수의 스마트 카메라를 클러스터링하여 협업하며 로드 밸런싱을 수행하는 알고리즘을 제안하고, Kubernetes 환경에서 시뮬레이션을 통해 여러 가지 상황에서 성능을 검증하여 엣지 컴퓨팅에서의 AI 연산을 보다 효율적으로 수행할 수 있는 방법을 제시한다.

1. 서론

AI 연산의 빠른 응답 시간을 제공하기 위해, AI 연산을 원거리의 클라우드 데이터센터가 아닌 서비스 요청 사용자와 근접한 엣지에서 수행되도록 하는 엣지 컴퓨팅이 각광받고 있다[1]. 스마트 카메라 네트워크도 연산 자원이 제한되어 엣지 컴퓨팅에 의존해 왔으나, 최근 스마트 카메라 엣지 클러스터를 구성해 다수의 스마트 카메라들이 협업하여 서비스를 처리하는 연구가 진행되고 있다[2]. 하지만 딥러닝에 사용되는 심층 신경망(DNN) 어플리케이션의 연산을 로컬에서 처리해야 할 경우 기존 클라우드 데이터센터에 비해 엣지 디바이스들의 에너지 및 메모리 용량 부족으로 인해 필요 연산 요구량을 만족시키지 못한다[3]. 이에 본 논문에서는 다수의 스마트 카메라들로 클러스터를 구성하여, 클러스터 내 스마트 카메라들 간의 협업을 통한 하이브리드 스케줄링 알고리즘(Dag task Hybrid Scheduling algorithm)을 제안한다. 제안 알고리즘 DHS 는 DNN 연산량과 스마트 카메라 부하 상태에 따라 서로 다른 휴리스틱 알고리즘들을 혼합 적용하여 DNN 추론 계산 처리 지연시간을 줄여준다. 시뮬레이션을 통해 로컬이나 엣지에서 DNN 추론에 필요한 계산을 모두 처리해줄 때에 비해 성능 향상이 있음을 확인하였다.

2. 제안 알고리즘

2.1 DNN 추론 계산 부하 구분

주어진 스마트 카메라에서의 연산량을 $\boldsymbol{w_i}$ 계산능력을 $\boldsymbol{C_i}$라고 했을 때 연산 처리 시간 $\boldsymbol{t_i}$는 $\boldsymbol{t_i = w_i/C_i}$로 계산할 수 있다. $\boldsymbol{t_i}$중 최솟값을 $\boldsymbol{t_{min}}$, 최댓값을 $\boldsymbol{t_{max}}$, 평균값을 $\boldsymbol{t_{mean}}$이라고 했을 때, 첫 번째 기준은 식 (1), (2)와 같이 결정된다.

$$t_{max} - t_{mean} > \delta_1 or\ t_{mean} - t_{min} > \delta_1 \rightarrow Unbalanced \quad (1)$$

$$t_{max} - t_{mean} \leq \delta_1\ and\ t_{mean} - t_{min} \leq \delta_1 \rightarrow Balanced \quad (2)$$

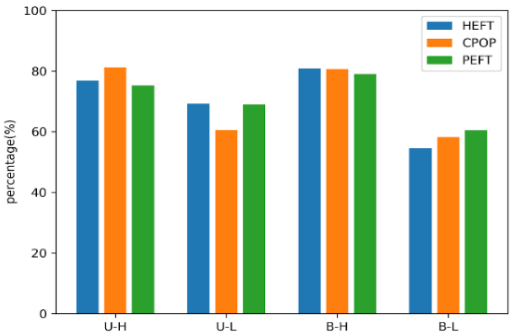
δ_1은 스마트 카메라의 성능과 DNN 어플리케이션의 연산 요구량에 따라 다른 한계점이다.

요청이 들어오는 DNN 어플리케이션의 개수를A, 스마트 카메라의 개수를 n이라 했을 때, 두 번째 기준은 식 (3), (4)와 같이 결정된다.

$$A/n \geq \delta_2 \rightarrow Heavy \quad (3)$$

$$A/n < \delta_2 \rightarrow Light \quad (4)$$

δ_2 역시 스마트 카메라의 성능과 DNN 어플리케이션의 연산 요구량에 따라 다른 한계점이다. 위 정의에 따라 주어지는 DNN 추론 계산 부하를 unbalanced 와 balanced 그리고 Heavy 와 Light 로 나누어 1)Unbalanced-Heavy(U-H), 2) Unbalanced-Light(U-L), 3) Balanced-Heavy(B-H), 4) Balanced-Light(B-L)의 각 4 가지 상황으로 구분한다.



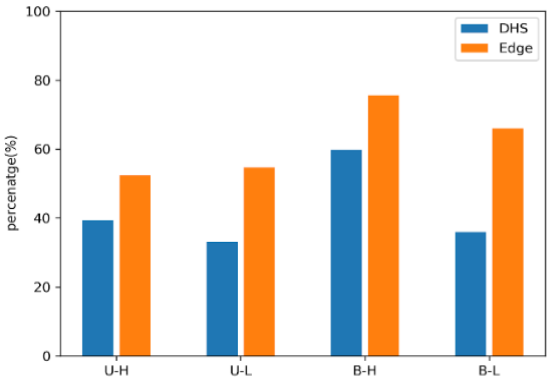
(그림 1) 부하 상황 별 알고리즘 비교

2.2. 이종 IoT DAG 태스크 스케줄링 알고리즘

DNN 어플리케이션은 Directed Acyclic Graph(DAG)로 표현할 수 있다. 서로 다른 연산 능력을 갖고 있는 엣지 디바이스들에게 DAG 태스크를 스케줄링하는 문제는 NP-hard 한 문제로 널리 알려져 있다. 따라서 본 논문에서는 DAG 태스크 스케줄링에 효과가 있다고 알려진 휴리스틱 알고리즘들인 Heterogeneous Earliest Finish Time(HEFT)[4], Critical Path On a Processor(CPOP)[4], Predict Earliest Finish Time(PEFT)[5] 알고리즘을 혼합하여 계산 처리 지연 시간을 최소로 만드는 하이브리드 스케줄링 알고리즘(DHS)을 제안한다. 그림 1 은 상황별로 HEFT, CPOP, PEFT 알고리즘의 연산 처리 지연 시간 비교를 보여준다. Heavy 상황에서는 PEFT, U-L 상황에서는 CPOP, B-L 상황에서는 HEFT 알고리즘이 최적임을 알 수 있다. 따라서 제안 알고리즘 DHS 에서는 1) Heavy 상황일 경우 PEFT 알고리즘을 사용해 DAG 태스크를 스케줄링 한다. 그렇지 않은 Light 상황일 경우 2) unbalanced 상황일 경우 CPOP 알고리즘을, balanced 상황일 경우 HEFT 알고리즘을 사용해 스케줄링 해준다.

3. 실험 및 성능 평가

본 논문에서 제안한 DHS 알고리즘의 성능을 평가하기 위해 4 개의 스마트 카메라와 하나의 엣지 서버로 클러스터를 구성하여 Kubernetes 환경[6]에서 실험했다. 스마트 카메라의 구현을 위해 하나의 Raspberry Pi 4B 와 세 대의 Jetson TX2 를, 3.6GHz CPU 클럭수와 24GB 의 메모리를 갖고 있는 엣지 서버 모델을 선정했다. Heavy 상황에서는 2 Alexnet, 2 GoogLeNet, 2 Resnet-50 request 를, Light 상황에서는 1 Alexnet, 1 GoogLeNet, 1 Resnet-50 request 를 발생시켰다. $\delta 1$ 은 0.1 로, $\delta 2$ 는 1 로 설정했다. 비교 알고리즘으로는 로컬 알고리즘과 엣지 알고리즘 두 가지를 상정한다. 요청되는 DNN 추론 연산을 모두 스마트 카메라 로컬에서 처리하는 알고리즘을 로컬 알고리즘, 엣지 서버로 보내 처리하는 방법을 엣지 알고리즘이라고 한다. 그림 2 는 각 상황에 따른 DHS 와 엣지 알고리즘의 처리 지연 시간을 로컬 알고리즘을 기준으로 한 백분율을 보여준다. U-H 부하에서 최소 12.9%, B-L 부하에서 최대 30.1%의 성능 향상을 확인할 수 있다.



(그림 2) DHS, 엣지 알고리즘 비교

4. 결론

본 논문에서는 DNN 연산을 처리하기 위해 클러스터 내 스마트 카메라 간의 협업을 통한 휴리스틱 로드 밸런싱 알고리즘을 적용해 소요 시간을 줄이는 기법을 제안하고 있다. 실험을 통해 전통적인 엣지 서버 기반 연산 처리보다 제안하는 알고리즘이 DNN 추론 계산 처리 지연 시간을 충분히 향상시킬 수 있음을 확인하였다.

사사문구

이 연구논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구결과임(No. 2022R1A2B5B01001683).

참고문헌

[1] E. Li, et al. "Edge AI: On-Demand Accelerating Deep Neural Network Inference via Edge Computing," IEEE Transactions on Wireless Communications, vol. 19, no. 1, pp. 447-457, 2020

[2] Zeng, Xiao, et al. "Distream: scaling live video analytics with workload-adaptive distributed edge intelligence." Proceedings of the 18th Conference on Embedded Networked Sensor Systems, New York, USA, 2020, pp. 409-421

[3] Li Zhou, et al. “Adaptive parallel execution of deep neural networks on heterogeneous edge devices.” Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. New York, USA. 2019. pp. 195–208.

[4] L. Jinhong, et al. "Learning to Optimize DAG Scheduling in Heterogeneous Environment," 23rd IEEE International Conference on Mobile Data Management, Paphos, Cyprus, 2022, pp. 137-146

[5] Middha, K., et al. “PEFT-Based Hybrid PSO for Scheduling Complex Applications in IoT”. Smart Computational Strategies: Theoretical and Practical Aspects, Singapore, Springer, 2019, pp. 137-146

[6] Y. Mao, et al. "Speculative Container Scheduling for Deep Learning Applications in a Kubernetes Cluster", IEEE Systems Journal, vol. 16, no. 3, pp. 3770-3781. 2022