심층신경망 어플리케이션을 처리하는 엣지 컴퓨팅 환경에서 강화학습을 이용한 Edge Orchestration

신지호, 한승재
연세대학교

jihoshin@yonsei.ac.kr, seungjaehan@yonsei.ac.kr

Edge Orchestration Using Deep Reinforcement Learning in Edge Computing Environments where Deep Neural Network Application be handled

Ji Ho Shin, Seung-Jae Han
Yonsei University

요 약

본 논문은 심층신경망을 이용한 어플리케이션을 서비스를 엣지 컴퓨팅 환경에서 지원하는 방법으로 심층신경망 모델의 특성을 이용하여 강화학습(Deep Reinforcement Learning)을 통한 Edge Orchestration 과정에 여러 모델을 서비스에 사용함으로써 어플리케이션의 처리량과 품질 간의 관계를 조절하는 방식을 제안한다. 다양한 구조의 심층신경망 벤치마크 데이터를 바탕으로 시뮬레이션을 진행하여 제안하는 방식의 성능을 확인하였다.

Ⅰ. 서 론

모바일 기기의 자원의 한계로 인하여 이용자들의 단말의 가까운 곳에 위치한 모바일 네트워크 엣지에서 컴퓨팅 능력과 서비스를 제공하는 Mobile Edge Computing (MEC)은 5G 시대의 핵심적인 요소로서 연구되어 왔다[1]. 최근 많은 분야에 적용되기 시작하는 심층신경망 기술이 적용된 어플리케이션 또한 이후에 MEC 환경에서 지원해야할 필요성이 높아지고 있다.

본 논문에서는 심층신경망 어플리케이션의 서비스를 효과적으로 지원하는 방법으로 강화학습을 이용한 Edge Orchestration 에 심층신경망의 모델적 특성을 반영한 서비스방식을 제안하고자 한다. 또한 시뮬레이션을 통해 제안한 방법이 심층신경망 어플리케이션 서비스를 제공 가능함을 보인다.

Ⅱ. 본 론

일반적으로 심층신경망 모델은 성능과 연산 요구량, 모델의 크기가 비례하는 경향을 보인다[2]. 본 논문에서는 클라우드 컴퓨팅 환경에 비해 가용자원이 제한적인 엣지 컴퓨팅 환경에서 심층신경망 모델을 사용하는 서비스에 이와 같은 특징을 이용하고자 한다. 가장 높은 성능을 보이며 많은 컴퓨팅 자원을 요구하는 하나의 모델만을 서비스에 사용하지 않고 여러 종류의 모델을 선택지로 가지며, Deep Reinforcement Learning (DRL) 에이전트가 현재 엣지 서버들의 상태를 반영하여 적합한 모델을 선택한 후 어플리케이션 요청을 Off-loading 하는 Edge Orchestration 방식을 통해 정확도와 같은 어플리케이션 품질관리와 서비스의 처리량을 조절한다.

본 논문의 방법을 시뮬레이션하기 위해 그림 1. 의 방식으로 엣지 시스템을 구성하였다. 엣지 서버들은 3 가지 종류로 구성되며 각 종류 별로 서로 다른 가용자원(MIPS, RAM)을 갖는다. 엣지 서버들은 각 종류 마다 2 개씩 총 6 개로 설정하였다. 단위 시간에 DRL 에이전트로 전달되는 어플리케이션 요청의 수는 Poisson 분포를 이용하여 개수를 설정하였다.

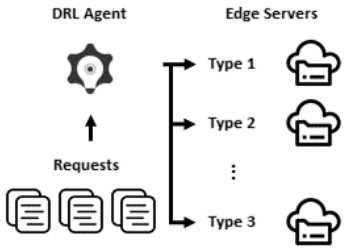

그림 1. Edge 시스템의 구조

생성된 요청은 먼저 DRL 에이전트에 전달되며 이후 에이전트는 개별 엣지 서버들의 현재 리소스 사용율과 엣지 서버의 대기열에 할당된 요청의 수를 고려하여 5 가지의 심층신경망 구조 중 하나를 선택하여 엣지 서버에 요청을 할당한다. 실험에서 사용한 심층신경망 모델의 크기와 컴퓨팅자원 요구 수치는 심층신경망 구조들의 벤치마크를[2] 기초로 설정하였으며 참고한 모델의 종류 연산요구량(G-FLOPS)와 모델의 정확도는 표 1. 의 내용과 같다. 사용한 DRL 에이전트는 OpenAI 에서 공개한 A2C (Advanced Actor Critic) [3]를 바탕으로 구현하였다. 엣지 서버에 전달된 요청은 FCFS (first come first served) 방식으로 처리된다. 학습의 목표는 시뮬레이션의 요청 처리결과 사용된 모델들의 평균 정확도가 크게 낮아지지 않으면서 처리한 요청의 수를 높이는 방향으로 설정하여 학습을 진행하였다.

종류	G-FLOPS	정확도
NASNET-A-Large	23.4	82.5
VGG-19	19.5	72.5
ResNet-101	7.5	77.2
MobileNet-v2	0.6	71.8
SqueezeNet-v1.1	0.4	58

표 1. 시뮬레이션에 사용한 심층신경망 모델 데이터

DRL 에이전트와의 성능을 비교하기 위한 다른 Orchestration 방식은 무작위로 선택을 하는 Random 방식과 순서대로 요청을 할당하는 Round Robin 방식을 사용하였다.

Ⅲ. 실험 결과 및 결 론

각 실험 조건에서 여러 번의 시뮬레이션을 진행한 후의 결과값의 평균을 표시하였다.

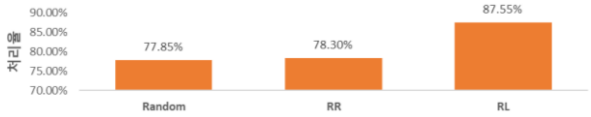

그림 2. 요청 처리율

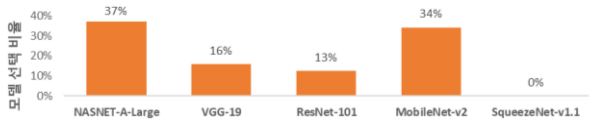

그림 3. DRL 에이전트가 모델을 선택한 비율

시뮬레이션 시간동안 어플리케이션 요청의 수가 800에서 6000 사이로 변화할 때 처리된 결과를 표시하였다. 실험결과 Random 방식과 Round Robin 방식은 전체 처리율에서 큰 차이를 보이지 않았으며 DRL 에이전트를 사용한 Orchestration 방식의 처리율이 더 높은 결과를 보였다. DRL 에이전트가 엣지 서버에 할당한 심층신경망 모델의 비율은 그림 3. 의 수치로 나타났다. 5 가지의 선택지로 주어진 모델 중 정확도가 다른 모델들에 비해 많이 낮았던 SqueezeNet 이 전체 처리율을 크게 낮추기 때문에 에이전트에 의해 선택된 비율이 없었다. 따라서 해당 모델이 포함된 Random 과 Round Robin 방식에 비해 서비스의 정확도 부분에 있어서도 4% 높은 결과를 얻을 수 있었다. 실험결과를 통해 심층신경망의 특성을 반영한 강화학습이 적용된 Edge Orchestration 이 서비스를 제공하기 위한 품질과 처리량을 학습을 통해 조절 가능함을 확인할 수 있었다.

본 논문에서는 엣지 컴퓨팅 환경에 심층신경망 기술을 사용하는 어플리케이션의 서비스를 지원하기 위해, 강화학습을 이용한 Edge Orchestration 방식에 심층신경망 모델의 특성을 이용한 방식을 제안하고, 벤치마크를 이용한 시뮬레이션을 수행하여 서비스 제공의 가능성을 확인하였으며 후속 연구로서 본 논문에서 실험한 단일 종류의 어플리케이션에서 더 나아가 서로 다른 종류로 구성된 다수의 어플리케이션 서비스 환경에서도 제안한 방식의 Orchestration 방법이 효과적으로 동작할 수 있는 방법을 연구할 계획이다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF2019R1H1A2079748)

참 고 문 헌

[1] 김상기, 박종대. (2016). 5G 를위한 MEC 기술동향, 전자통신동향분석, Vol. 31, No. 1, pp. 25-35. P-ISSN: 1225-6455

[2] Bianco, Simone, et al. "Benchmark analysis of representative deep neural network architectures." IEEE Access 6 (2018): 64270-64277.

[3] OpenAI Baseline code, A2C, Github, https://github.com/openai/baselines