

마이크로서비스 기반의 클라우드 엣지 AI 추론 서비스 개발 및 연구

서지현¹, 장수민¹, 차재근¹, 최현화¹, 김대원¹, 김선욱¹,

¹ 한국전자통신연구원

blue_reverb@etri.re.kr, jsm@etri.re.kr, jgcha@etri.re.kr,
hyunwha@etri.re.kr, won22@etri.re.kr, swkim99@etri.re.kr

Development and Study of Cloud-Edge AI Inference Service Based on Microservices

Ji-Hyun Seo¹, Su-min Jang¹, Jae-geun Cha¹, Hyun-hwa Choi¹, Dae-won Kim¹, Sun-wook Kim¹,

¹Electronics and Telecommunications Research Institute (ETRI)

요 약

최근 딥러닝을 이용한 영상 분석은 자율주행, 감시카메라 등 다양한 서비스에 필수적으로 활용되고 있으며 실시간 처리 및 보안 요소를 만족하기 위해 기존의 클라우드 컴퓨팅 방식의 단점을 개선한 클라우드 엣지 컴퓨팅 방식을 적용하는 사례가 크게 증가하고 있다. 하지만 사용자 및 단말과 가까운 위치에서 딥러닝 추론을 진행하는 클라우드 엣지 서버는 클라우드 서버와 비교하여 컴퓨팅 자원이 충분하지 않을 경우가 많으며 기존의 딥러닝 모델을 그대로 클라우드 엣지 환경에 적용하는 것은 자원 활용 측면에서 여러가지 문제점들을 갖고 있다. 따라서 본 논문에서는 마이크로서비스 구조를 통해 자원을 보다 유연하게 활용할 수 있도록 개선된 딥러닝 모델로 대규모의 클라이언트 요청을 처리 가능한 동영상 데이터 추론 서비스인 G-Edge AI 추론 서비스 개발에 대해 설명한다.

1. 서론

자율주행차와 같은 이동체부터 도로, 도시 등의 CCTV로 획득하는 대용량 영상 분석에 이르기까지 현대사회에서 다양한 동영상 데이터 분석 수요는 날이 급증하고 있다. [1-2] 동영상 분석 분야에서 딥러닝은 실시간 요구조건을 만족하며 높은 성능을 보이기 때문에 대부분의 서비스에서 필수적으로 채택하고 있다. 그러나 서비스를 제공하는 과정에서 대용량의 동영상 데이터를 전송하고 처리하기 위해 많은 양의 컴퓨팅 자원이 들 뿐 더러, 높은 QoS(Quality of Service)를 요구했기에 이에 클라우드 컴퓨팅(Cloud Computing) 방식이 대두되었다. 그러나 클라우드 컴퓨팅이란 중앙집중식 구조로 이루어져 있기 때문에 다수의 단말에서 빠른 속도로 증가하는 데이터를 중앙 서버로 전송하는 데 정체 현상이 발생할 수 있었고, 처리 결과를 얻기까지 오랜 시간이 걸릴 수 있었다. 이 과정에서 해결책으로 여겨진 방식은 클라우드 엣지 컴퓨팅(Cloud-edge Computing) 방식이다. 이 방식은 데이터 입력 단말과 가까운 거리에 엣지 서버가 위치해 있어 데이터를 전송하기 위한 네트워크 대역폭 및 지연시

간에 대한 부담을 줄일 수 있었다. 또한 단말과 가까운 서버로부터 빠른 처리 결과를 얻을 수 있어 클라우드 컴퓨팅의 단점을 개선하였다.

그러나 여전히 기존의 딥러닝 모델로 클라우드 엣지 서버의 컴퓨팅 자원을 효율적으로 사용하기에는 어려움이 존재한다. 클라우드 서비스는 딥러닝 모델의 성능을 보장하기 위해 중앙 서버에 충분한 컴퓨팅 자원(고성능 GPU 등)을 장착한 경우를 기본으로 한다. 반면, 컴퓨팅 자원이 충분하지 않은 경우가 많은 클라우드 엣지 서버 환경에서 무겁고 연산량이 많은 기존의 딥러닝 추론 모델을 이용하기에는 제약이 따른다. 이는 클라우드 컴퓨팅 환경의 딥러닝 모델과 클라우드 엣지 컴퓨팅 모델 간 차이를 두어야 함을 시사하며, 클라우드 엣지 서버에서 딥러닝 서비스를 제공하기 위해서는 유연한 자원 활용에 더욱 초점을 맞출 수 있는 방안이 필요하다는 것을 의미한다. 따라서 본 논문에서는 마이크로서비스 구조를 딥러닝 모델에 적용하여 클라우드 엣지 컴퓨팅 환경에서 유연한 자원 활용이 가능하도록 개선한 동영상 데이터 추론 서비스 개발 내용에 대해 설명하고자 한다.

<표 1> G-Edge AI 서비스 작업 데이터 정보

Key	의미
#	DB 내의 등록된 번호
User	Request 를 보낸 사용자를 구분하기 위한 식별자
Path	동영상이 저장소에 업로드 된 경로
Type	동영상 파일의 확장자
Status	8 개의 상태(Enqueue, Preprocessing, Preprocessing_complete, Inferencing, Inferencing_complete, Postprocessing, Postprocessing_complete, Exit)로 현 작업의 상태 정보를 의미
Start Time	작업의 처리가 시작된 시간
End Time	작업의 처리가 완료된 시간
Iteration	오류로 인한 작업 재시작 횟수, 일정 횟수 이상이면 작업 실패로 간주

4. 결론 및 향후 연구

본 논문에서는 클라우드 엣지 컴퓨팅 환경에서 동영상 데이터를 빠르고 유연하게 추론할 수 있도록 마이크로서비스 구조를 통해 자원 활용 능력을 개선한 G-Edge AI 추론 서비스 개발에 대해 설명하였다. 이를 통해 사용자로부터 가까운 위치에서 동영상 데이터들을 수집 및 처리함으로써 데이터 전송의 지연 및 작업시간을 최소화할 수 있다. 또한, 엣지 간 유연하고 신속한 자원 사용에 특화된 딥러닝 모델을 통해 확장성이 높고 오류에 강건한 추론 서비스를 사용자에게 제공할 수 있을 것으로 예상된다.

향후 연구로는 실 서비스 환경에서도 무리없이 서비스 제공이 가능할 수준으로 서비스를 확장해 나갈 계획이다. 서비스 메시지를 구성하여 더욱 다양한 상황에서도 서비스 간 원활한 통신을 유지하도록 개선할 예정이며, 현재 클러스터에 작업을 완료하기 위한 적당한 자원이 존재하지 않는 경우에는 가용 자원을 능동적으로 추가할 수 있는 시스템을 개발할 예정이다. 또한 효율적인 자원 활용을 할 수 있는 딥러닝 추론 모델뿐만 아니라 학습 모델까지도 확장하여 클라우드 엣지 컴퓨팅 환경의 딥러닝 서비스를 더욱 확장해 나가고자 한다.

Acknowledgement

본 연구는 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행되었음 (2020-0-00116, 10msec 미만의 서비스 응답 속도를 보장하는 초저지연 지능형 클라우드 엣지 SW 플랫폼 핵심 기술 개발)

참고문헌

- [1] R. Nayak, P. Umesh, and D. Santos. "A comprehensive review on deep learning-based methods for video anomaly detection." *Image and Vision Computing* 106, 2021.
- [2] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang. "New generation deep learning for video object detection: A survey." *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [3] Kubernetes, <https://kubernetes.io/docs/home/>
- [4] A. Bochkovskiy, W. Chien-Yao, and L. Hong-Yuan. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934*, 2020.