마이크로서비스 기반의 클라우드 엣지 AI 추론 서비스 개발 및 연구

서지현 ¹, 장수민 ¹, 차재근 ¹, 최현화 ¹, 김대원 ¹, 김선욱 ¹, ¹한국전자통신연구원

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 단말과가까운 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 마이크로서비스구조를 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 CCTV 로 획득하는 대용량 영상 분석에 이르기까지 현대사회에서 다양한 동영상 데이터 분석 수요는 나 날이 급증하고 있다. [1-2] 동영상 분석 분야에서 딥러 닝은 실시간 요구조건을 만족하며 높은 성능을 보이 기 때문에 대부분의 서비스에서 필수적으로 채택하고 있다. 그러나 서비스를 제공하는 과정에서 대용량의 동영상 데이터를 전송하고 처리하기 위해 많은 양의 컴퓨팅 자원이 들 뿐 더러, 높은 OoS(Quality of Service) 를 요구했기에 이에 클라우드 컴퓨팅(Cloud Computing) 방식이 대두되었다. 그러나 클라우드 컴퓨팅이란 중 앙집중식 구조로 이루어져 있기 때문에 다수의 단말 에서 빠른 속도로 증가하는 데이터를 중앙 서버로 전 송하는 데 정체 현상이 발생할 수 있었고, 처리 결과 를 얻기까지 오랜 시간이 걸릴 수 있었다. 이 과정에 서 해결책으로 여겨진 방식은 클라우드 엣지 컴퓨팅 (Cloud-edge Computing) 방식이다. 이 방식은 데이터 입력 단말과 가까운 거리에 에지 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 엣 지 서버의 컴퓨팅 자원을 효율적으로 사용하기에는 어려움이 존재한다. 클라우드 서비스는 딥러닝 모델 의 성능을 보장하기 위해 중앙 서버에 충분한 컴퓨팅 자원(고성능 GPU 등)을 장착한 경우를 기본으로 한다. 반면, 컴퓨팅 자원이 충분하지 않은 경우가 많은 클 라우드 엣지 서버 환경에서 무겁고 연산량이 많은 기 존의 딥러닝 추론 모델을 이용하기에는 제약이 따른 다. 이는 클라우드 컴퓨팅 환경의 딥러닝 모델과 클 라우드 엣지 컴퓨팅 모델 간 차이를 두어야 함을 시 사하며, 클라우드 엣지 서버에서 딥러닝 서비스를 제 공하기 위해서는 유연한 자원 활용에 더욱 초점을 맞 출 수 있는 방안이 필요하다는 것을 의미한다. 따라 서 본 논문에서는 마이크로서비스 구조를 딥러닝 모 델에 적용하여 클라우드 엣지 컴퓨팅 환경에서 유연 한 자원 활용이 가능하도록 개선한 동영상 데이터 추 론 서비스 개발 내용에 대해 설명하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서 제안하는 클라우드 엣지 AI 추론 서비스에 있어 핵심 개념들에 대해 간단히 소개하고, 3 장에서 AI 추론 서비스의 전체적인 구성 및 동작 과정을 설명한다. 그리고 마지막으로 본 서비스를 통해 기대할 수 있는 효과 및 향후 연구 방향에 대하여 논의하며 마무리한다.

2. 관련 연구

2-1. 클라우드 엣지 컴퓨팅(Cloud-Edge Computing)

클라우드 컴퓨팅은 최근까지도 빅데이터 개념과 결합하여 4차 산업혁명의 핵심 기술로 대두되어 왔다. 그러나 IoT 기기들이 기하급수적으로 증가함에 따라 단말과 서버의 데이터 통신량 폭증으로 그 한계점을 드러내고 있다. 이에 대한 해결 방안으로 등장한 클 라우드 엣지 컴퓨팅은 분산 클라우드 형태의 컴퓨팅 방식의 일종으로 클라우드 컴퓨팅의 한계점으로 거론 되는 처리 지연, 데이터 전송 지연, 대역폭 폭증 등에 대한 문제를 개선하였다. 중앙 서버에서 데이터 처리 를 지원하는 기존의 클라우드 컴퓨팅 방식과는 달리 데이터를 입력으로 받는 단말 또는 서비스 사용자와 지리적으로 가까운 위치에서 클라우드 컴퓨팅 기술을 제공한다. 또한 기존의 클라우드 컴퓨팅이 클라우드 와 단말 간의 수직적 협업만을 제공했다면, 클라우드 엣지 컴퓨팅은 엣지와 엣지 간의 수평적 협업도 고려 한다. 서비스, 데이터, 자원에 대한 엣지 간 협업을 통해 유연한 서비스 제공이 가능해진다. 본 논문에서 컨테이너 기반 오케스트레이션 Kubernetes [3]를 기반으로 서비스 전반의 클러스터 구 성 및 개선한 AI 추론 서비스 배포를 진행하였다.

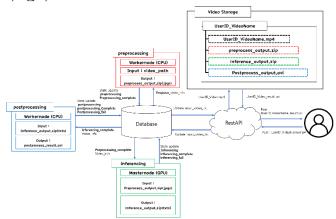
2-2. 마이크로서비스 구조

마이크로서비스 구조(Microservices Architecture, MSA)란, 서비스 지향 구조(Service Oriented Architecture, SOA)로부터 파생된 소프트웨어 아키텍처로 독립된 서비스들의 집합을 통해 하나의 어플리케이션을 구축 하는 방식을 의미한다. 마이크로서비스 구조의 장점 으로는 첫번째로 기능 단위로 서비스를 분할하기 때 문에 배포 단위가 작으며, 두번째로 서비스들이 독립 적으로 동작하기 때문에 자동화에 유리하다는 장점이 있다. 세번째로 확장성이 좋은 유연한 서비스를 구성 할 수 있으며, 마지막으로 개발자는 오류에 강건한 어플리케이션 개발을 할 수 있다는 장점이 있다. 서 비스 지향 구조는 상대적으로 단단하게 결합되어 스 토리지 버스라고 하는 전문 소프트웨어를 통해 통신 하지만 각 서비스는 캡슐화되어 비교적 가벼운 프로 토콜(예로, HTTP, REST)을 통해 통신한다. 본 논문에서 는 영상 분석을 위한 비전 딥러닝 모델을 사용하는

컴퓨팅 자원에 따라 그룹화하여 마이크로서비스 구조를 적용했다. 그 후, 작은 서비스 단위의 기능들이 원활하고 협력적으로 작동할 수 있도록 조율하는 과정을 적용하였다.

3. G-Edge AI 추론 서비스

클라우드 엣지 컴퓨팅 환경에서 동작하는 마이크로 서비스 기반 AI 추론 서비스의 전체적인 구조는 그림 1과 같다.



(그림 1) G-Edge AI 추론 서비스 구조도

영상 추론을 위한 딥러닝 모델은 크게 세 부분으로 나누어질 수 있다: 이미지의 입력 크기를 조절하거나 더 나은 추론 성능을 위한 영상처리를 진행하는 전처리 과정, 딥러닝 네트워크를 통해 객체의 특징을 추출하여 인식하고, 객체의 위치를 특정하는 추론 과정, 추론 과정에서 나온 결과를 이미지에 표시하거나 부가적인 영상처리를 진행하는 후처리 과정. 이 중, 전처리 과정과 후처리 과정은 컴퓨팅 자원 중 CPU 를 주로 활용하는 과정이며, 추론 과정은 GPU 를 이용하여 많은 연산량을 빠르게 처리하는 과정이다. 따라서 엣지 간 유연한 컴퓨팅 자원 활용을 극대화하고자 본논문에서는 마이크로서비스 구조를 활용하여 기존의 딥러닝 모델(YoloV4 [4])을 세 부분으로 나누었으며, 각 과정은 독립적으로 동작할 수 있도록 컨테이너 기반의 Pod로 배포되었다.

또한 컨테이너 기반의 오케스트레이션 플랫폼인 Kubernetes 를 활용하여 클라우드 엣지 기반의 환경을 생성할 때 고려해야 할 점은 각 서비스 간의 통신이충돌없이 원활하게 이루어져야 한다는 점이다. 이를 위해 동영상 데이터가 업로드 됨에 따라 표 1과 같은 형태로 작업의 정보를 데이터베이스에 저장하고, 서비스 Pod 들이 데이터베이스와 통신하며 작업 상태를 공유하도록 하였다.

<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 | 8개의 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. 이를통해 사용자로부터 가까운 위치에서 동영상 데이터들을 수집 및 처리함으로써 데이터 전송의 지연 및 작업시간을 최소화할 수 있다. 또한, 엣지 간 유연하고신속한 자원 사용에 특화된 딥러닝 모델을 통해 확장성이 높고 오류에 강건한 추론 서비스를 사용자에게제공할 수 있을 것으로 예상된다.

향후 연구로는 실 서비스 환경에서도 무리없이 서비스 제공이 가능할 수준으로 서비스를 확장해 나갈계획이다. 서비스 메시를 구성하여 더욱 다양한 상황에서도 서비스 간 원활한 통신을 유지하도록 개선할예정이며, 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 추론모델뿐만 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 응답속도를 보장하는 초저지연 지능형 클라우드 엣지 SW 플랫폼핵심 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.