서버리스 플랫폼에서 GPU 지원 및 인공지능 모델 추론에 적합한 함수 구조에 관한 연구

황동현*, 김동민, 최영윤, 한승호, 전기만, 손재기
*전자부품연구원 휴먼 IT 융합연구센터
e-mail : {dhhwang89*, dmkim*, cyy03576*, tmdgh7186*, kmjeon*, jgson*} @keti.re.kr

A Study on Function which supported GPU and Function Structure Optimization for AI Inference

Dong-Hyun Hwang*, Dongmin Kim, Young-Yoon Choi,
Seung-Ho Han, Gi-Man Jeon, Jae-Gi Son
*Dept. of Human IT Convergence Research Center, Korea Electronics Technology

요 약

서버리스 프레임워크(Serverless Framework)는 마이크로서비스 아키텍처의 이론을 클라우드와 컨테이너를 기반으로 구현한 것으로 아마존의 AWS(Amazon Web Service)와 같은 퍼블릭 클라우드 플랫폼이 서비스됨에 따라 활용도 높아지고 있다. 하지만 현재까지의 플랫폼들은 GPU 와 같은 하드웨어의 의존성을 가진 인공지능 모델의 서비스에는 지원이 부족하다. 이에 본 논문에서는 컨테이너 기반의 오픈소스 서버리스 플랫폼을 대상으로 엔비디어-도커와 k8s-device-plugin 을 적용하여 GPU 활용이 가능한 서버리스 플랫폼을 구현하였다. 또한 인공지능 모델이 컨테이너에서 구동될 때 반복되는 가중치 로드를 줄이기 위한 구조를 제안한다. 본 논문에서 구현된 서버리스 플랫폼은 객체 검출 모델인 SSD(Single Shot Multibox Detector) 모델을 이용하여 성능 비교 실험을 진행하였으며, 그 결과 인공지능 모델이 적용된 서버리스 플랫폼의 함수 응답 시간이 개선되었음을 확인하였다

1. 서론

클라우드 컴퓨팅이 대중화된 이후 많은 회사가 자사의 웹 애플리케이션을 퍼블릭 클라우드 벤더로 배포하고 있다. 모노리틱 구조의 애플리케이션을 클라우드 컴퓨팅에 배포하고자 하는 경우 자동 확장, 지속적 배포 및 통합, 내고장성이라는 측면에서 한계점이 있다[1]. 이러한 한계점을 극복하기 위해 애플리케이션 구조를 서버리스 컴퓨팅으로 변경하는 방법이 주목받고 있다.

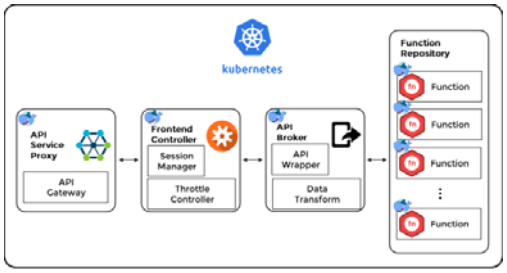
2012 년 이후 높은 성능으로 인해 인공지능에 대한 관심이 커졌다. 인공지능을 적용한 시스템은 하드웨어 의존성, 소프트웨어 의존성, GPU 사용이 가능한 배포 시스템의 부재로 인해 개발 및 배포가 어려웠다[2-3].

2. 관련 연구

2-1. 도커(Docker)와 쿠버네티스(Kubernetes)

도커를 이용하여 개발자는 개발 환경을 응용프로그램 관리 방식과 유사한 인프라로 관리할 수 있다.

쿠버네티스와 도커의 조합은 인공지능 모델을 배포하기 위한 서버리스 프레임워크를 구현할 때 좋은 도구로써 사용된다. 그림 1 은 쿠버네티스와 도커 기반으로 구현된 서버리스 플랫폼 구조를 보여준다.



(그림 1) 쿠버네티스를 이용한 컨테이너 기반의 서버리스 플랫폼 구조

2-2. 엔비디아-도커(NVIDIA-Docker)

도커 컨테이너에서 엔비디아 그래픽카드를 사용하게 될 경우 하드웨어 및 관련 소프트웨어에 종속된다.

이 문제를 해결하기 위해서 엔비디아는 도커에서 하드웨어와 소프트웨어에 약한 의존성을 갖는 엔비디아 컨테이너 런타임을 개발하였다.

2-3. k8s-device-plugin

k8s-device-plugin 은 각 클러스터에 장착되어있는 GPU 의 개수 노출, GPU 상태 추적, 쿠버네티스 클러스터 안에서 컨테이너가 GPU 를 사용할 수 있게 허용하는 기능을 제공한다

3. 인공지능 모델을 위해 GPU 를 지원하는 서버리스 플랫폼

3-1. GPU supported Serverless Framework

GPU 함수를 지원하기 위하여 k8s-device-plugin 을 쿠버네티스의 데몬셋으로 실행하였다. 그 결과 쿠버네티스는 클러스터 내부의 GPU 정보와 GPU 자원을 스케쥴링할 수 있게 되었다.

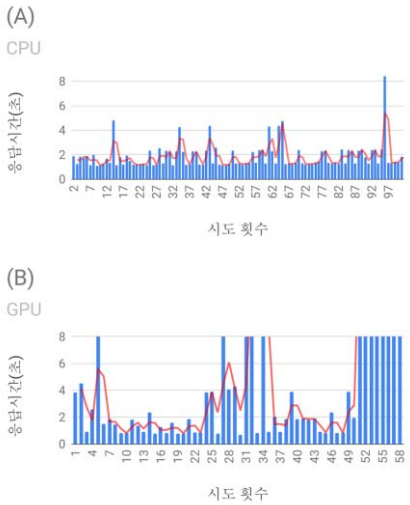
3-2. Callable Object

서버리스의 무상태성은 인공지능 모델 추론 시 일회성으로 호출되는 모델 객체 선언, 가중치 로드를 반복적으로 실행하도록 하며 이러한 반복적 실행은 응답 시간을 지연시킨다. 이를 해결하기 위하여 파이썬의 호출 가능한 객체(Callable Object)를 이용하는 방식으로 함수 구조를 변경하였다.

4. 결과

구조가 변경된 서버리스 플랫폼의 성능은 인공지능 모델 중 객체 검출 모델인 SSD 의 TensorFlow 구현체를 이용하여 추론 시간을 측정하였다. 실험은 100 회의 요청에 대한 응답 시간을 측정하였고 응답 시간들의 표준편차를 계산하였다.

그림 3 은 구조 변경 전의 서버리스 플랫폼에서 CPU 함수의 응답 시간을 측정한 것이다. 막대그래프는 개별 요청에 대한 응답 시간을, 선 그래프는 함수 응답 시간의 추세를 보여준다. 구조 변경 전의 CPU 함수의 응답 시간은 평균 1.87 초, 표준편차는 1.062 이었다. 구조 변경 전의 GPU 함수의 응답 시간은 평균 56.55 초, 표준편차 150.710 였다. 구조 변경 후의 CPU 함수의 응답 시간은 평균 0.26 초, 표준편차는 0.012 였다. GPU 함수 구조 변경 후의 GPU 함수의 응답 시간은 평균 0.02 초, 표준편차 0.001 이었다. 서버리스 플랫폼의 함수 구조 변경 전, 후 각 함수 별 평균 응답 시간과 표준편차는 표 3 에서 확인할 수 있다.



(그림 3) 함수 구조 변경 전 서버리스 플랫폼의 함수 응답시간

<표 3> 서버리스 플랫폼 함수 구조 변경 전/후 함수 응답 시간과 표준편차

성능 \ 구조		변경 전	변경 후
CPU	평균 (초)	1.87	0.26
	표준 편차	1.062	0.012
GPU	평균 (초)	56.55	0.02
	표준 편차	150.710	0.001

실험 결과 구조가 변경된 서버리스 플랫폼의 함수 응답 시간이 전반적으로 상당히 빨라진 것을 확인할 수가 있었으며, 응답 시간의 편차 또한 적은 것을 확인할 수가 있었다.

5. 결론

본 논문에서는 오픈소스 서버리스 플랫폼에서 엔비디아-도커, k8s-device-plugin, 함수 구조의 변경을 통해서 GPU 지원을 가능하게 하였다. 또한 이러한 함수 구조의 변경의 성능을 측정하기 위하여 인공지능 모델 중 객체 검출 모델인 SSD 를 통해서 함수 구조별 응답 지연시간을 측정하여 변경된 서버리스 플랫폼이 인공지능 모델을 배포하기 더 적합한 플랫폼이라는 것을 검증하였다.

Acknowledgements

이 논문은 2017 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2017-0-00255, (지능정보-총괄/1 세부) 자율지능 디지털 동반자 프레임워크 및 응용 연구개발)

참고문헌

[1] Mario Villamizar, Oscar Garcés, Harold Castro, Mauricio Verano, Lorena Salamanca, Rubby, Casallas, Santiago Gil “Evaluating the monolithic and the microservice architecture pattern to deploy web applications in the cloud” 2015 10th Computing Colombian Conference (10CCC), At Bogotá, Colombia

[2] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, Thomas Zimmermann “Software Engineering for Machine Learning: A Case Study” ICSE-SEIP ‘10 Proceeding of the 41st International Conference on Software Engineering: Software Engineering in Practice Pages 291-300

[3] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Franc¸ois Crespo, Dan Denniso.“Hidden Technical Debt in Machine Learning Systems” Advances in Neural Information Processing Systems 28(NIPS 2015)