클라우드 플랫폼에서의 딥러닝 기반 웹 어플리케이션 서비스 성능 비교 분석

김주찬, 범정현, 추현승
성균관대학교 소프트웨어대학
wncks0928@g.skku.edu, bumjh@skku.edu, choo@skku.edu

Performance Comparison Analysis of Deep Learning-based Web Application Services on Cloud Platforms

Ju-Chan Kim, Junghyun Bum, Hyun-Seung Choo
College of Software, Sungkyunkwan University

요 약

최근 코로나바이러스감염증-19(COVID-19)가 확산됨에 따라 화상회의, 온라인 게임, 스트리밍 등과 같은 다양한 온라인 서비스들의 트래픽이 크게 증가하면서 원활한 서비스 제공을 위한 서버 자원 관리의 중요성이 강조되고 있다. 이에 따라 서버 자원을 전문적으로 관리해주는 클라우드 서비스의 수요도 증가하는 추세이다. 하지만 대다수의 국내 기업들은 성능의 불확실성, 보안, 정서적 이질감 등을 이유로 클라우드 서비스 도입에 어려움을 겪고 있다. 따라서 본 논문에서는 클라우드 서비스의 성능의 불확실성을 해소하기 위해 클라우드 시장 BIG3 기업(아마존, 마이크로소프트, 구글)의 클라우드 서비스의 성능을 비교하였다.

1. 서론

클라우드 컴퓨팅 서비스 시장은 코로나바이러스감염증-19(COVID-19)로 인해 회의, 수업 또는 민원 접수 등의 행위가 비대면 방식으로 전환하게 되며 온라인 서비스 수요가 크게 증가했다. 이에 따라 온라인 서비스를 제공하는 서버의 자원 관리에 대한 중요성이 강조되고 있다. 하지만 기업 또는 기관의 입장에서 늘어나는 인터넷 수요에 대한 서버 확충은 관리 인력, 공간, 비용 등의 증가로 현실적인 어려움이 많다. 이로 인해 클라우드 컴퓨팅 시스템의 도입이 꾸준히 증가하고 있지만 여전히 성능의 불확실성, 보안, 정서적 이질감 등의 이유로 인해 저해 받고 있다[1]. 본 논문에서는 클라우드 서비스 시장의 BIG3(아마존의 AWS, 마이크로소프트의 Azure, 구글의 GCP)의 성능 비교를 통해 클라우드 컴퓨팅 시스템 도입을 저해하는 성능의 불확실성을 해소하고자 한다. 특히, 클라우드 플랫폼을 이용하는 어플리케이션 서비스 제공자 입장에서 서비스 성능을 비교하기 위해 클라우드 플랫폼을 통해 가상머신을 할당받아 딥러닝 어플리케이션 서비스를 배포하고 서비스 응답시간 및 비용을 비교하고자 한다.

2. 관련 연구

클라우드 서비스는 기능에 따라 SaaS(Software as a Service), PaaS(Platform as a Service), IaaS(Infrastructure as a Service)로 분류된다. SaaS는 클라우드를 통해 어플리케이션 소프트웨어를 제공하며 대부분의 SaaS는 어플리케이션 또는 웹 브라우저를 통해 직접 실행되어 클라이언트 측에서 소프트웨어를 직접 다운하거나 설치하는 행위가 필요없다. PaaS는 사용자 정의 응용 프로그램을 개발하는데 필요한 운영체제 및 미들웨어, 런타임 등을 포함하는 플랫폼을 가상화해 제공하고, 개발자는 서비스 내에서 어플리케이션 개발에 집중할 수 있다. IaaS는 클라우드를 통해 네트워크, 서버, 운영체제 및 스토리지를 가상화해 제공하며 필요에 따라 물리적 컴퓨팅 자원의 변경 및 교체가 필요한 경우 기존 방식보다 빠르고 효율적인 대응이 가능하다.

국내에서 클라우드 컴퓨팅 시스템 도입을 저해하는 요인으로 꼽히는 성능의 불확실성은 IaaS의 형태로 제공되는 클라우드 서비스의 가용성과 성능에 대한 신뢰성 검증이 어렵기 때문에 발생한다. 위 연구에서는 이러한 기능적, 성능적 불확실성 해소를 위

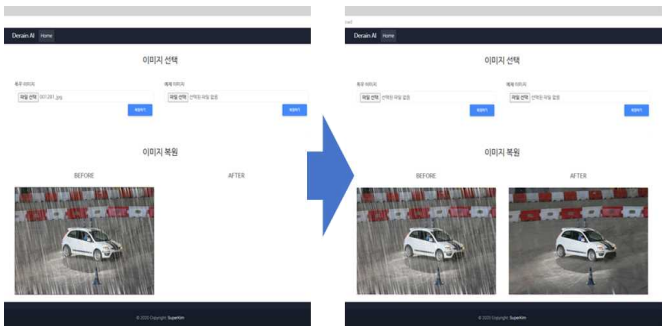
해 공개 소프트웨어를 기반으로 IaaS 클라우드 컴퓨팅 서비스에 대한 성능시험 기법[2]이 제안되었다.

3. 클라우드 플랫폼에서의 딥러닝 어플리케이션 서비스 성능 비교 분석

3.1 상용 클라우드 플랫폼

아마존의 AWS(Amazon Web Service)는 클라우드 서비스 시장의 선두 주자로 클라우드 시장의 초창기부터 오랜 기간 축적된 데이터와 인프라를 바탕으로 넓은 가용영역에서 안정적인 서비스 제공한다. 마이크로소프트의 Azure는 2014년 클라우드 서비스 시장에 후발주자로 뛰어들어 높은 자본력과 기술력을 바탕으로 무서운 성장세를 보이고 있으며, 특히 2019년 이후로 마이크로소프트가 가진 Windows 및 SQL 서버 라이선스를 아웃소싱 대상에서 제외하면서 경쟁 플랫폼에서 마이크로소프트 제품 활용도를 낮춰 Azure 서비스의 경쟁력을 높였다. 구글의 GCP(Google Cloud Platform)는 IaaS 서비스에서 할당 가능한 가상머신에 대한 커스터마이징이 앞서 소개한 기법들과 비교해 자유로우며, 구글에서만 제공되는 TPU(Tensor Processing Unit)라는 딥러닝을 위한 텐서 전용 연산 장치를 제공하고 있다.

본 연구에서는 각각의 플랫폼에서 제공하는 IaaS를 활용하여 어플리케이션 서비스를 배포하고, 이용자 수에 따른 응답 속도 및 재화 비용을 비교한다.

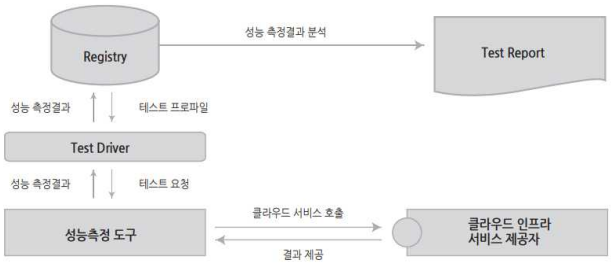


(그림 1) 딥러닝 기반 웹 어플리케이션 서비스 접속 화면. 사용자로부터 임의의 영상을 입력 받는 화면(위), 빗줄기를 제거하고 반환된 결과 웹 화면(아래)

3.2 딥러닝 기반 웹 어플리케이션 서비스

클라우드 서비스의 공통적 기능을 통해 딥러닝 기반 웹 어플리케이션을 제공하는 방법은 두 가지가 있다. 첫 번째는 클라우드 플랫폼이 제공하는 PaaS를 통해 웹 어플리케이션 배포해 서버를 호스팅 하는 방법이고, 두 번째는 IaaS를 통해 가상머신을 할당받아 직접 서버를 호스팅 하는 방법이다. 본 연구에서는 컴퓨팅 성능을 모니터링하기 위해 IaaS 방식을 채택하여 실험을 진행하였다.

본 연구에서 딥러닝 서비스를 제공하는 어플리케이션 시스템은 다음과 같다. 구현을 위해 사용된 언어는 Python이며 서버 호스팅과 딥러닝을 위해 Flask 모듈과 Pytorch 모듈을 사용하였고, 제공되는 서비스는 사용자가 입력한 이미지에서 빗줄기를 제거하여 보여주는 서비스이다. 빗줄기 제거를 위해 학습된 ResNet-18 모델을 사용하였다[3]. 해당 시스템의 동작 과정은 다음 그림 1과 같다. 첫째, 사용자로부터 임의의 이미지를 입력받는다. 둘째, 딥러닝 모델은 입력된 이미지로부터 결과를 추정한 뒤, 이미지를 스토리지에 저장하고 결과를 사용자에게 웹 페이지 형태로 반환한다.



(그림 2) 클라우드 성능 측정 테스트 흐름도.

<표 1> 클라우드 가상머신 하드웨어 및 운영체제 사양

Platform	CPU	RAM	Storage	Operating System
GCP	Intel(R) Xeon Platinum 8272CL CPU	16GB	50GB	Windows DataBase Server 2019
AWS	Intel(R) Xeon Platinum 8259CL CPU	16GB	50GB	Windows DataBase Server 2019
Azure	Intel(R) Xeon CPU	16GB	50GB	Windows DataBase Server 2019

3.3 성능 측정 환경

클라우드 서비스의 경우 실제 웹을 통해 시스템 성능 측정이 가능하고, 시간에 따라 네트워크 변화가 존재하기 때문에 측정한 정보의 객관성 및 신뢰성 확보를 위해 지속적인 성능 측정이 가능한 에이전트(Agent) 방식을 채택하였으며, 전체 테스트 흐름도는 그림 2와 같다. Test Driver가 성능 측정 도구를 통해 클라우드 서비스를 호출하면, 클라우드 인프라 서비스 제공자는 서비스에 대한 결과를 제공하고, 성능 측정 도구는 응답 결과에 대해 성능 측정 결과를 Test Driver에게 반환한다. 성능 측정 결

<표 2> Locust를 이용한 성능 평가 결과

Platform	Requests	Average (ms)	Min (ms)	Max (ms)	RPS	50%ile (ms)	90%ile (ms)	100%ile (ms)
GCP	97.42	40,651.67	1,828.7	76,145.81	0.80	18,179	24,795	30,792
AWS	104.61	40,698.48	**1,768.4**	73,847.89	0.78	17,953	24,601	30,653
Azure	101.66	**39,254.50**	1,824.0	**65,894.58**	0.76	**17,761**	**24,506**	**30,540**

<표 3> 가상 머신 사용 요금(7일)

Platform	1	2	3	4	5	6	7	계(원)
GCP	3,766	4,602	4,304	4,727	4,563	4,292	3,701	29,955
AWS	3,840	4,250	4,111	4,884	4,807	4,707	3,804	30,403
Azure	3,680	3,987	4,506	3,708	4,200	5,206	3,456	**28,743**

과는 Registry에 저장되며, 그 결과를 분석해 Test Report에 저장하게 된다. 테스트는 Locust 성능 측정 도구를 사용해 6시간 간격으로 7일간 총 28회 시행하였으며, 공정한 성능 측정을 위해 각각의 클라우드 플랫폼에서 책정된 요금제를 기준으로 동일한 금액으로 할당 가능한 가상머신을 선정하였다. 각각의 플랫폼에서 사용한 하드웨어와 운영체제의 사양은 다음 표 1과 같다.

3.4 실험 결과

IaaS를 통한 딥러닝 기반 어플리케이션 서비스 제공에 대한 네트워크 성능 측정 결과는 다음 표 2와 같다. Request는 총 요청 수, Average는 평균 응답 속도(ms), Min은 최소 응답 속도, Max는 최대 응답 속도, RPS는 초당 요청 수, n% ile은 n백분위에 대한 응답 속도로, n은 최대 클라이언트 수 대 현재 클라이언트 수의 비이다. 최대 클라이언트 수는 1,000으로 설정하였다. 7일간 테스트 결과의 평균으로 봤을 때 1) Azure, 2) AWS, 3) GCP 순으로 응답 속도 및 처리 시간이 빠른 것으로 나타났으며, 특히 최대 응답 속도를 살펴봤을 때 Azure가 가장 빠른 응답 속도를 보였다. 또한 해당 실험에서 총 7일에 걸쳐 소요된 비용은 표 3과 같다.

4. 결론 및 향후 연구 계획

본 연구에서는 딥러닝 기반 어플리케이션 서비스를 각각의 클라우드 클랫폼 Iaas를 통해 제공하고 해당 서비스의 네트워크 성능과 호스팅 비용을 비교하였다. 그 결과 Azure가 할당된 가상머신의 하드웨어 성능에 비해 근소하게 높은 성능을 보였다. 이는 Azure가 실험에 사용된 Windows 운영체제에 최적화되어 있기 때문이라고 분석된다. 또한, 표 2에서 전체적인 평균 응답 속도가 매우 낮은 것을 확인할 수 있다. 이는 CPU를 사용해 딥러닝 서비스를 제공하면서 오버헤드가 발생했기 때문이다.

향후 연구에서는 위 실험을 GPU 환경으로 확장해서 보다 안정적인 결과를 얻고자 한다. 또한 각 플랫폼 간의 기능적 차이를 명시하여 사용 목적에 따른 분석도 진행하고자 한다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성(Grand ICT연구센터) 사업(IITP-2021-2015-0-00742), 과학기술정보통신부 및 정보통신기획평가원의 ICT명품인재양성 사업(IITP-2021-2020-0-01821), 2021년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원(NRF-2020R1A2C2008447, 딥적대적러닝 기반의 버추얼 엣지: 자가감독형 엣지 이동성, 리소스 배치 및 할당)의 연구결과로 수행되었음.

참고문헌

[1] D.-H. Kim and J.-H. Lee, and Y-P Park. "A study of factors affecting the adoption of cloud computing" The Jounal of Society for e-Business Studies, 17(1). 111-136. 2012.

[2] K.-J. You and D.-S. Go, "Study on the performance test technique of open SW-based cloud computing", The Journal of Korean Institute of Information Technology, 10(7), 185-192, 2012.

[3] J.-C. Kim and C.-H. Son, "Structure-Aware Residual Network for Rain Streaks Removal", The Journal of Korean Institute of Information Technology, 18(10), 87-100, 2020.