|  |  |
| --- | --- |
| Performance and Cost Comparison of Cloud Services for Deep Learning Workload | |
| **저자** | Dheeraj Chahal, Mayank Mishra, Surya Palepu, Rekha Singhal |
| **학회(연도)** | ICPE(International Conference on Performance Engineering)(2021) |
| **한줄 요약** | |
| 클라우드 서비스를 이용한 딥러닝 워크로드에 대한 성능 비교 | |
| **요약** | |
| 본 논문에서는 클라우드 서비스인 aws사의 Sagemaker와 severless플랫폼인 Lambda를 비교하여 Sagemaker의 지연시간이 짧은 점과, lambda가 비용 절감 효과 및 트래픽 급증에 적합하다는 각각의 강점을 언급하며, 해당하는 두가지의 딥러닝 워크로드에 대해 논의하였다.  aws SageMaker와 Lambda를 사용하여 각각의 딥러닝 워크로드의 성능을 비교한다. NISER를 사용한 세션 기반 추천시스템을 소개하였고 이러한 추천 시스템을 클라우드로 마이크레이션 하는 방법론에 대해 주로 논의하였다. 연구는 SageMaker 엔드포인트에서 사용되는 다른 인스턴스의 응답 시간과 비용을 평가하며 CPU및 GPU인스턴스를 비교하였다. 또한 온프레미스 배포의 응답시간을 비교하였고 위의 다양한 내용에 해당하는 성능 및 비용 비교를 제시하며 lambda에 해당하는 서버리스 아키텍처가 버스트 워크로드에 대해 더 나은 성능을 제공한다는 결론을 도출하였다. 또한 ML플랫폼이 장기 실행 워크로드에 더 적합하다고 강조하였다. | |