실시간 빅데이터 기반 딥러닝 모델 추론 시스템

박경석*, **, 유찬희*, **, 김유선*, ** 엄정호*
*한국과학기술정보연구원
**UST 빅데이터과학과
gspark@kisti.re.kr

An Inference System for Deep Learning Model Based on Real-time Big Data

Kyongseok Park*, **, Chan Hee Yu*, **, Yuseon Kim*, ** Jung-Ho Um*
*Korea Institute of Science and Technology Information
**Department of Big Data Science, UST

요 약

최근의 빅데이터 처리 환경은 실시간 빅데이터를 기반으로 하고 있다. 실시간 빅데이터 처리를 위해서는 기존의 배치처리 방식의 빅데이터 기술에서 발생하는 기술적 요구를 포함하여 추가적으로 요구되는 다양한 문제들을 고려해야 한다. 기계학습 모형을 활용한 의사결정 지원 시스템의 경우 모형 개발을 위한 배치처리 기술과 함께 모형의 배포와 최적화 등도 고려되어야 하며 발전 설비나 제조, 공정, 배송 등의 분야에서 발생하는 대규모 실시간 데이터를 이용하여 추론을 수행해야 한다. 본 연구에서는 센서 데이터를 활용한 예측 모형 개발과 실시간 데이터 처리 그리고 추론을 위한 모델 배포와 최적화 과정을 지원하는 시스템 환경을 제공하여 실제 현장에서 발생하고 있는 데이터를 활용하여 실증을 수행하였다.

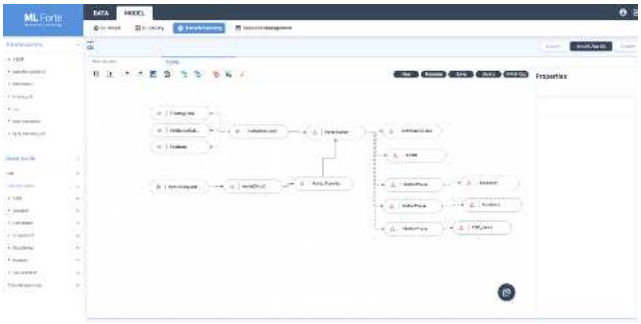
1. 서론

각종 센서의 증가와 더불어 이들 센서에서 생성되는 데이터를 비롯하여 연관된 데이터 역시 증가하고 있다. 특히 센서에서 생성되는 데이터는 배치처리와 실시간 처리에 대한 요구가 동시에 존재한다. 전통적인 빅데이터는 주로 배치처리를 위한 기술에 의존해 왔으나 현재 상당수의 빅데이터는 실시간 처리 필요성이 요구된다.

대규모 센서 데이터를 이용하여 예측을 수행할 경우 최근 기계학습 기반의 모형 활용이 증가하고 있다. 특히 딥러닝과 같은 기계학습 모형은 개발에도 많은 시간과 비용이 소요되지만 개발한 모형을 현장에 적용하여 서비스로 운영하기 위해서는 추가적인 절차와 최적화 작업이 필요하다. 일반적으로 모형을 개발하는 분석가들은 이러한 절차에 제대로 준비되지 않았기 때문에 이를 효율적으로 수행하기 위해서는 엔지니어링적인 지원이 필요하다[1].

본 연구에서는 대규모 발전 설비를 비롯한 제조, 생산 분야에서 수집되는 데이터를 이용하여 기계학습 모형을 개발하고 모델 서버에 배포하여 실시간 처리를 통한 모델 추론과정을 효율화할 수 있는 시스템을 제안하고 있다. 이를 통해 의사결정 지원 시스템에 적용하여 실제 데이터 수집과 처리 환경에서 예측을 수행한 사례를 살펴보고자 한다.

2. 제안 시스템 및 적용 결과



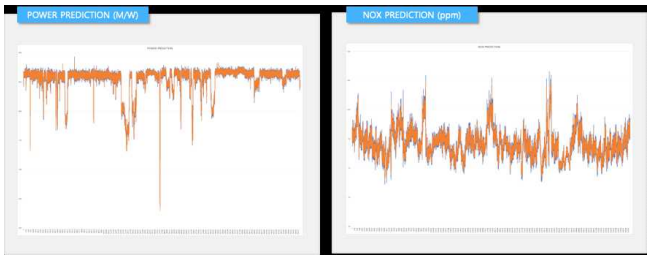
(그림 1) 실시간 모델 처리 환경

전통적인 빅데이터 분석은 주로 파일시스템이나 데이터베이스 등에 저장된 역사적 데이터(historical data)를 이용하여 집계 및 요약을 하거나 모형을 개발하는 배치처리 방식으로 이루어진다. 이러한 배치처리 방식은 여전히 많은 분야에서 활용되고 있으며 중요한 빅데이터 분석 방식으로 자리매김하고 있다. 그러나 센서 정보를 활용해야 하는 응용 분야의 경우 대규모 배치처리 못지않게 실시간 빅데이터 처리 기술 역시 매우 중요하다. 실시간 빅데이터의 경우 배치처리를 위한 응용보다 훨씬 복잡한 처리 절차와 기

술을 요구하게 된다. 배치처리의 경우 과거에 축적된 데이터를 추출하여 원하는 형식으로 가공하고 분석을 하면 된다. 당연히 이러한 처리 절차에도 많은 노력과 기술적 도전이 필요하다. 실시간 빅데이터의 경우는 실시간으로 데이터가 변경되고 있기 때문에 실시간으로 입력되는 데이터에 대한 필터링(filtering), 조인(join), 윈도우 처리(windowing), 버퍼링(buffering) 등 고려해야 하는 요소가 훨씬 많다.

이러한 환경에서 기계학습 모형을 적용할 경우 다수의 모형을 동시에 처리하고 의사결정을 하기 위한 추론 단계에서 이러한 요소들에 대한 고려가 충분히 이루어져야 한다. 다수의 기계학습 모형이 같은 데이터를 입력으로 받아 서로 다른 처리 절차를 거쳐 추론을 할 수도 있고 서로 다른 데이터를 입력으로 받아 서로 다른 모형에서 각각 추론을 할 수도 있다. 특히 발전 설비와 같은 경우 초당 수십만에서 수백만 개 이상의 센서에 대한 정보를 받아 추론을 수행해야 하기 때문에 의사결정 지원 시스템에 적용하기 위한 기술적 난이도가 높아지게 된다.

본 연구에서는 대규모 센서 데이터와 다수의 기계학습 모형에 대한 실시간 추론을 위해 MMS(Multi Model Server) 기반의 추론 시스템을 적용하여 수십만 개의 센서 데이터를 이용하여 수십 개 또는 수백 개의 모형을 동시에 처리할 수 있는 환경을 제공하고 있다[2, 3]. 모형 개발자들이 모형 개발 환경에서 자신의 모형을 개발한 후 바로 모델 서버로 배포하고 서비스를 제공할 수 있도록 하여 모형 개발과 서비스 적용 간의 시간을 단축하고 기술적 괴리를 제거함에 따라 모형 개발자들이 온전히 자신의 모형에 고민하고 집중할 수 있도록 하였다[4].



(그림 2) 모델 서버를 적용한 추론 결과

위 그림은 발전소에서 생성되는 실시간 센서 데이터를 이용하여 딥러닝 모형을 개발한 후 모델 서버에 배포하여 발전 용량과 NOx 배출량을 실시간으로 예측한 사례이다. 두 모형의 경우 대부분 서로 다른 유형의 센서를 사용하여 예측을 수행하고 있으며 시계열 예측을 위한 LSTM 계열의 모형을 적용하여 예측을 수행하였다.

3. 결론

발전 설비나 제조 현장에서 생성되는 데이터는 대부분 센서를 통해 생성되는 데이터이다. 이러한 센서 데이터를 분석하기 위해서는 실시간 처리 기술이 고려되어야 한다. 또한 다수의 모델을 동시에 처리하고 추론하기 위한 기술과 개발한 모형을 모델 서버로 간결하게 배포할 수 있는 환경이 필요하다. 본 연구에서 제안한 시스템을 통해 모형 개발과 배포 환경 간의 간극을 줄이고 대용량의 센서 데이터를 실시간으로 처리할 수 있도록 지원함에 따라 다양한 응용 환경에 활용될 것으로 기대한다.

사사

이 논문은 산업통상자원부의 재원으로 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구입니다. (No. 20181110100420)

이 논문은 중소벤처기업부의 재원으로 중소기업기술정보진흥원(TIPA)의 지원을 받아 수행한 연구입니다. (S3126610)

참고문헌

[1] Mu Li, David G. Andersen, Jun Woo Park, Alexander J. Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J. Shekita, and Bor-Yiing Su, “Scaling Distributed Machine Learning with the Parameter Server”, OSDI, 2014, pp. 583-598

[2] Tianqi Chen and Thierry Moreau, “TVM: An Automated End-to-End Optimizing Compiler for Deep Learning”, OSDI, 2018, pp. 579-594

[3] Eggers, S. J., Emer, J. S., Levy, H. M., Lo, J. L., Stamm, R. L., and Tullsen, D. M., “Simultaneous multithreading: a platform for next-generation processors”, IEEE Micro, 2017, 12-19

[4] 박경석, 유찬희, Komal Sarda, 엄정호, “분산 딥러닝 모델 개발을 위한 고수준 분석 플랫폼”, 한국정보처리학회 추계학술대회, 2020, pp. 804-806