Apache Spark와 OpenCV를 활용한 분산 클러스터 컴퓨팅 환경 대용량 이미지 머신러닝 시스템

김하윤[1], 김원집[2], 이협건[3], 김영운[4]
[1]숭실사이버대학교 ICT공학과
[2]한국폴리텍대학 서울강서캠퍼스 빅데이터과
[3]한국폴리텍대학 서울강서캠퍼스 빅데이터과
[4]한국폴리텍대학 서울강서캠퍼스 빅데이터과
let_hykim@naver.com, kbg5174@naver.com,
hglee67@kopo.ac.kr, luckkim@kopo.ac.kr

Image Machine Learning System using Apache Spark and OpenCV on Distributed Cluster

Hayoon Kim[1], Wonjib Kim[2], Hyeopgeon Lee[3], Young Woon Kim[4]
[1]Dept. of ICT Engineering, Korea Soongsil Cyber University
[2]Dept. of Big Data, Seoul Gangseo Campus of Korea Polytechnic
[3]Dept. of Big Data, Seoul Gangseo Campus of Korea Polytechnic
[4]Dept. of Big Data, Seoul Gangseo Campus of Korea Polytechnic

요 약

성장하는 빅 데이터 시장과 빅 데이터 수의 기하급수적인 증가는 기존 컴퓨팅 환경에서 데이터 처리의 어려움을 야기한다. 특히 이미지 데이터 처리 속도는 데이터양이 많을수록 현저하게 느려진다. 이에 본 논문에서는 Apache Spark와 OpenCV를 활용한 분산 클러스터 컴퓨팅 환경의 대용량 이미지 머신러닝 시스템을 제안한다. 제안하는 시스템은 Apache Spark를 통해 분산 클러스터를 구성하며, OpenCV의 이미지 처리 알고리즘과 Spark MLlib의 머신러닝 알고리즘을 활용하여 작업을 수행한다. 제안하는 시스템을 통해 본 논문은 대용량 이미지 데이터 처리 및 머신러닝 작업 속도 향상 방법을 제시한다.

1. 서론

국내·외 빅 데이터 시장은 지속해 성장하고 있으며, 이에 따라 처리할 빅 데이터는 기하급수적으로 증가하고 있다.[1] 이는 기존 컴퓨팅 환경에서 데이터 처리의 어려움을 야기한다. 특히 이미지 데이터 처리 속도는 데이터양이 많을수록 현저하게 느려진다.

클러스터 컴퓨팅은 여러 대의 컴퓨터를 하나의 시스템으로 연결하여 작업을 처리하는 방식이며 대용량 데이터 처리 솔루션으로서 주목받고 있다. 이러한 분산 클러스터 컴퓨팅 기술은 빅데이터 처리 분야에서 빠르게 발전하고 있으며, 이를 위한 다양한 프레임워크들이 등장하고 있다.

그중 Apache Spark는 범용 분산 클러스터 컴퓨팅 프레임워크로서, 인메모리 기반의 대용량 데이터 고속 처리가 가능한 오픈 소스 엔진이다.

본 논문은 대용량 이미지 데이터 처리 속도 지연 문제의 해결을 위해 분산 클러스터 컴퓨팅 환경의 대용량 이미지 머신러닝 시스템을 제안한다. 제안하는 시스템은 분산 클러스터 컴퓨팅 프레임워크인 Apache Spark와 컴퓨터 비전 프로그래밍 라이브러리인 OpenCV를 통해 구현된다.

본 논문의 구성은 다음과 같다. 2장은 제안하는 시스템 구성의 핵심 요소들을 소개하고, 3장은 제안하는 이미지 머신러닝 시스템에 대해 설명한다. 4장은 제안하는 시스템의 활용 방안과 향후 과제에 대한 고찰로 결론을 맺는다.

2. 관련 연구

본 장은 제안하는 시스템의 핵심 구성 요소인 Apache Spark와 HDFS를 제시한다.

2.1. Apache Spark

Apache Spark(이하 Spark)는 오픈 소스 분산 클러스터 컴퓨팅 프레임워크이다. Spark는 인메모리 기반의 빠른 데이터 처리 속도를 지원하며, Spark MLlib을 통해 다양한 머신러닝 알고리즘을 제공한다. 또한 Spark는 클러스터 확장에 용이한 아키텍처

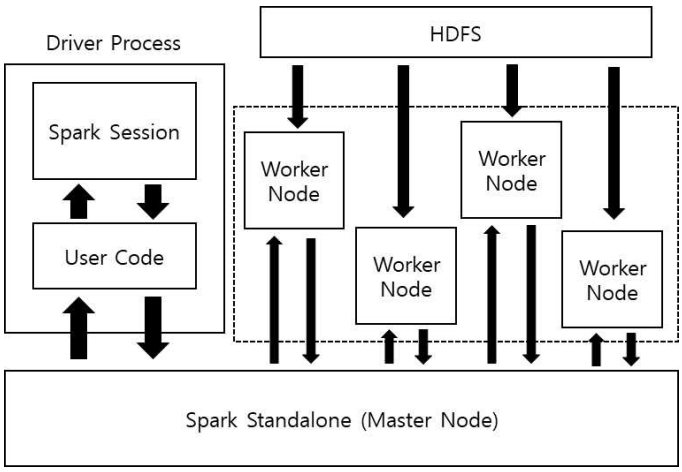
를 제공하여 단일 소스 코드를 수천 개 노드 내결함성 클러스터로 확장한다.

2.2. HDFS(Hadoop Distributed File System)

HDFS는 클러스터에서 대규모 데이터를 저장하고 처리하기 위한 분산 파일 시스템이다. HDFS는 분산 컴퓨팅 프레임워크인 Apache Hadoop에 의해 제공되며 대용량 데이터의 저장과 처리를 위한 고가용성, 확장성, 안정성을 제공한다.

3. 제안

본 논문은 Apache Spark와 OpenCV를 활용한 분산 클러스터 컴퓨팅 환경 대용량 이미지 머신러닝 시스템을 제안한다. 제안하는 시스템은 분산 클러스터 환경에서 구현되며, 대용량 이미지 데이터 처리 및 머신러닝 속도 향상을 위한 방법을 제시한다. [그림 1]은 제안하는 시스템 구성도를 나타낸다.



[그림 1] 시스템 구성도

제안하는 시스템은 Driver Process, Spark Master Node, Spark Worker Node, HDFS로 구성된다.

Driver Process는 Spark Application의 메인 진입점으로서, spark-submit을 통해 전송받은 User Code를 실행시킨다. Spark Session은 전송받은 작업 명령을 Spark Standalone(Master Node, 클러스터 관리자)에게 전달한다.

User Code는 OpenCV를 통해 구현되는 필터링, 분할 등 이미지 처리 작업과 Spark MLlib이 제공하는 머신러닝 알고리즘이 적용될 수 있다. 사용자는 작업을 이미지 데이터 특성에 맞게 Gradient-Boosted Trees, Random Forest 등의 알고리즘을 선택하여 머신러닝 알고리즘을 제출할 수 있다.

Spark Standalone으로 구현된 Master Node는 작업을 클러스터 전역으로 분배하여 Spark Worker Node들에 분배하여 실행시킨다. Master Node는 클러스터의 작업 실행에 필요한 자원을 관리한다.

HDFS는 분산 파일 시스템으로서 요청받은 작업을 수행하기 위해 이미지 데이터를 저장한다. HDFS는 데이터를 작은 블록들로 분할하고 여러 노드에 분산 저장하여 Worker Node들의 분산 리소스 접근을 허용한다.

Spark Worker Node는 HDFS에서 이미지 데이터를 Spark Image DataSet으로 변환한다. 변환된 Spark Image DataSet은 클러스터에 분산 저장되며 각 Node들에서 병렬 처리될 수 있다. 각 Node들은 Master Node로부터 받은 User Code의 OpenCV, MLlib 알고리즘을 통해 이미지 분류, 객체 탐지, 세그멘테이션 등의 머신러닝 프로세스를 수행한다.

객체 탐지/세그멘테이션/이미지 분류 모델은 Spark MLlib을 활용하여 학습된 머신러닝 모델로서 이미지 분류, 객체 탐지, 세그멘테이션 작업을 수행한다.

제안하는 시스템을 통해 사용자는 대용량 이미지 데이터 처리 및 머신러닝 알고리즘을 구현하고 분산 클러스터 시스템에서의 실행을 명령할 수 있다.

4. 결론

본 논문은 대용량 이미지 데이터 처리 및 머신러닝을 위한 분산 클러스터 컴퓨팅 환경의 대용량 이미지 머신러닝 시스템을 제안한다. 제안하는 시스템은 분산 컴퓨팅 프레임워크 Apache Spark와 컴퓨터 비전 라이브러리 OpenCV를 통해 구현된다.

제안하는 시스템을 통해 본 논문은 대용량 이미지 데이터 처리 및 머신러닝 작업 속도 향상 방법을 제시한다.

참고문헌

[1] JangYeol Lee, ChoonSung Nam and DongRyeol Shin, "Machine Learning Bigdata Education Platform using Apache Spark", Korea Computer Congress, Jeju, Korea, 2017,1531.