양자화 시뮬레이션 플랫폼의 GUI-기반 통합을 위한 관리 모듈

임채민*, 조상영**, 임승호**
한국외국어대학교 *언론정보전공. **컴퓨터공학부
chaemin.lim.hufs@gmail.com, sycho@hufs.ac.kr, slim@hufs.ac.kr

A Management Module for GUI-based Integration of Quantization Simulation Platform

Chaemin Lim*, Sang-Young Cho**, Seung-Ho Lim**
*Dept. of Journalism & Information, **Division of Computer Engineering
Hankuk University of Foreign Studies

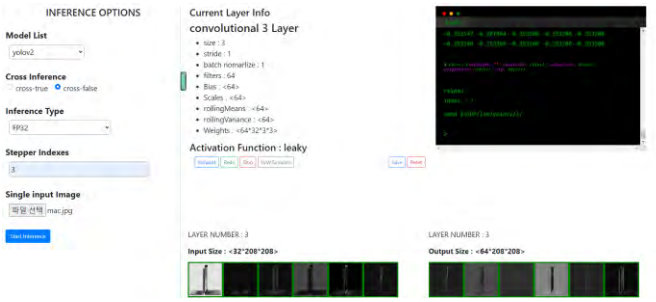
요 약

신경망 모델 데이터의 양자화는 모델 크기를 줄이고 추론 시간을 단축할 수 있다. 본 논문에서는 양자화를 지원하는 시뮬레이션 플랫폼의 전체 동작 관리를 위한 관리 모듈에 대해 기술한다. 시뮬레이션 플랫폼이 다중 사용자를 지원하고 다양한 기능을 지원하기 때문에 효율적인 관리 모듈의 구현은 중요하다. 관리 모듈은 웹 서비스의 후위단으로 설계되었으며 라우터와 이벤트 수신자, 프로세스 관리자, 파일 관리자, 세션 관리자, 보조 기능 등을 구현하였다.

1. 서론

인공 신경망(Artificial Neural Network)은 다양한 응용 분야에서 성공적으로 적용되어 지대한 관심을 받고 있다. 많은 산업계와 학계에서 인공 신경망의 세부적인 연구를 수행하고 있다. 인공 신경망이 발전하면서 네트워크 모델이 점점 복잡해지고 있다. 레이어의 수가 많아지고 모델의 파라메터의 수가 사람이 다루기에 어려운 상황에 도달하였다. 인공 신경망 연구를 용이하게 하기 위한 모델 및 데이터의 시각화 연구에도 많은 투자를 하고 있다. Google TensorFlow 의 TensorBoard[1]와 NVIDIA 에서 개발한 NVIDIA Digits[2]가 대표적 사례이다.

본 논문에서는 합성곱 신경망(Convolutional Neural Network: CNN)의 양자화 시뮬레이션을 위한 GUI 기반 통합 환경의 관리 모듈 구현에 대해 다룬다. 그림 1 은 양자화 플랫폼에서 CNN 모델의 계층 단위 정보를 시각화 하는 GUI 의 기능을 보여준다. 양자화 시뮬레이션 환경은 CNN 객체 검출 알고리즘인 YOLO[3]와 그 프레임워크인 Darknet 을 기반으로 구축되었다. Darknet 의 실수 추론 기능에 반정수 하이브리드 추론 기능과 양자화를 이용한 정수 추론 기능이 추가되어 양자화 성능을 검사할 수 있는 플랫폼이다. 양자화 시뮬레이션 플랫폼에서 수행되는 추론 동작을 검증하기 위한 신경망 모델 및 추론의 시각화 및 분석 도구를 GUI 형태로 통합하였으며 이를 지원하기 위한 관리 모듈을 제안하고 구현한다.

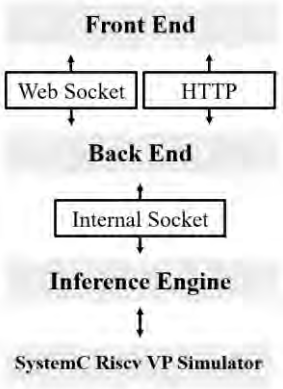


(그림 1) 양자화 플랫폼 GUI 의 추론 시각화 기능.

2. 본론

신경망 모델 분석 및 추론 확인을 위한 GUI 인터페이스는 신경망 모델의 양자화 파라메터 값의 수정을 위한 시각화와 추론 동작 분석 기능을 제공한다. 추가로 신경망 추론 모델의 시각화, 다중 사용자를 위한 추론 엔진(Inference Engine) 구동, 추론 디버깅, 신경망 모델 편집, 추론 과정 시각화, 데이터 증가(Augmentation), 파일 관리 기능이 있다.

전체 양자화 시뮬레이션 플랫폼의 동작은 웹 서비스로 구현된 GUI 환경에서 수행할 수 있기 때문에 다중 사용자의 동시 원격 사용이 가능하다. 본 논문에서 다루는 관리 모듈은 웹 서비스의 전위단(Front-end)과 추론 엔진의 중간에서 동작하며 웹 서비스 구현 기준으로 후위단(Back-end)에 해당한다.



(그림 2) 시뮬레이션 플랫폼에서 후위단의 통신 구조.

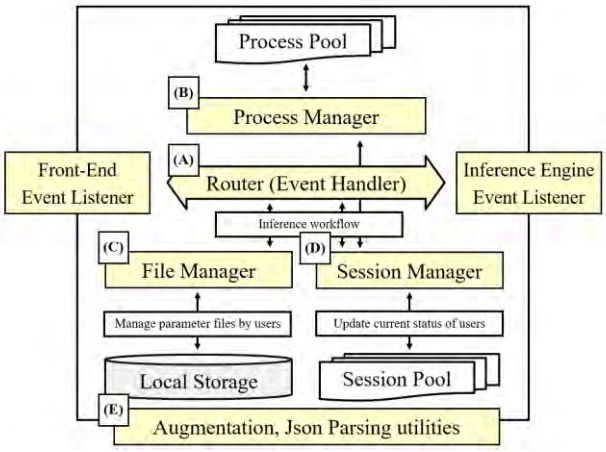
시뮬레이션 플랫폼의 추론 엔진은 Darknet 기반의 신경망 연산 엔진으로서 실수 연산과 양자화 모델 연산, 교차 추론 연산 등의 기능을 지원한다. 후위단은 추론 엔진 기반 통합 시뮬레이션 플랫폼의 서버 환경을 구성하며, 추론 엔진의 동작을 웹에서 확인할 수 있게 한다. 후위단은 전위단과는 웹 프로토콜을 이용한 통신 방법을 사용하고, 추론 엔진과는 내부 소켓을 통한 통신 방법을 이용한다. 후위단은 전위단과 추론 엔진 사이에서 추론 과정의 파라미터로 이용되는 데이터를 접근하여 관리하며 전위단과 추론 엔진을 연결한다. 그림 2 은 시뮬레이션 플랫폼의 전체적 통신 구조를 보여준다.

(그림 3) 시뮬레이션 플랫폼의 성능 평가 도구.

플랫폼은 양자화된 모델의 평가를 위해 수치화된 데이터와 시각화된 신경망을 제공한다. 플랫폼은 양자화된 신경망 모델의 추론 과정에서 이용되는 매개 변수 값의 통계적 정보를 제공한다. 추론 과정에서는 계층 단위의 입출력 데이터를 시각화하여 가중치에 의해 활성화되는 부분을 확인할 수 있다. 또한 신경망 모델의 전체적인 구조와 커널의 크기, 활성 함수의 종류와 같은 매개 변수의 정보를 확인할 수 있게 한다 (그림 3).

2.1 후위단의 구조

후위단은 Node.js[4]와 Express.js[5] 엔진을 이용하며 크게 다섯개의 모듈로 구성된다: 라우터와 이벤트 수신자(Router and Event Listener: REL), 프로세스 관리자(Process Manager), 파일 관리자(File Manager), 세션 관리자(Session Manager), 보조 기능(Utilities) (그림 4). REL 이 후위단의 주요 모듈로 주로 전위단의 요청에 대한 응답을 하는 서비스 루틴의 기능을 수행한다. 나머지 네 모듈은 REL 이 수행하는 응답 루틴을 보조한다.



(그림 4) 후위단 내부 구조.

A. 라우터와 이벤트 수신자 (REL)

REL 은 후위단의 관리자로써 주요 요청 처리 기능을 담당한다. 파일 관리, 추론, 가중치 수정, 이미지 증대, 이미지 시각화 등 플랫폼의 모든 기능을 웹 브라우저에 시각적으로 표현하기 때문에 후위단은 다양한 요청에 따라 이벤트 서비스 루틴을 수행한다. 대용량 파일의 효율적인 전송을 위하여 웹 소켓과 HTTP 프로토콜을 혼용한다.

이벤트 수신자는 동시에 동작하는 전위단과 추론 엔진을 위한 두 개의 수신자가 있다. 전위단 이벤트 수신자는 전위단의 HTTP 와 웹 소켓 요청을 처리한다. 웹 소켓을 통해서 대용량 파일의 업로드와 다운로드를 스트림을 통해 처리할 수 있다. 또한, 사용자의 요청 없이도 서버에서 바로 사용자에게 웹 소켓 메시지를 보낼 수 있다.

추론 엔진 이벤트 수신자는 서버 내부에서 동작하며 소켓 통신을 통해 추론 엔진과 상호작용한다. 사용자에 의하여 실행된 추론 엔진의 결과 파일을 수신하여 파일 저장 공간에 저장한다. 사용자 단위로 파

일 저장 공간을 가지고 있기 때문에 추론 엔진은 다중 사용자간 독립적으로 동작한다.

B. 프로세스 관리자

추론 엔진은 독립적인 프로그램으로 개발되었다. 따라서 Node.js 런타임에서 추론 엔진을 실행시킬 때에는 후위단 프로세스의 자식 프로세스로 실행시킨다. 프로세스 관리자는 세션을 참조하여 추론 엔진 프로세스를 검사하며 추론 엔진 프로세스의 비정상적 종료 또는 오동작이 발생하면 이에 대한 후처리를 수행한다.

C. 파일 관리자

파일 관리자는 다수의 사용자가 동시에 추론 엔진을 사용할 수 있도록 사용자 단위로 파일 디렉토리를 관리하며 파일 디렉토리에는 사용자가 추론 수행을 위해 전위단에서 업로드한 네트워크 모델 파일과 가중치 파일이 저장되며 사용자의 가중치 파일 수정과 파일 업/다운로드의 작업이 가능하다. 또한 추론 엔진은 추론 과정에서 이 파일들의 경로를 파라미터로 전달받아 동작하기 때문에 추론에 필요한 파일을 생성할 때 이 파일 저장 구조를 따라야 한다.

D. 세션 관리자

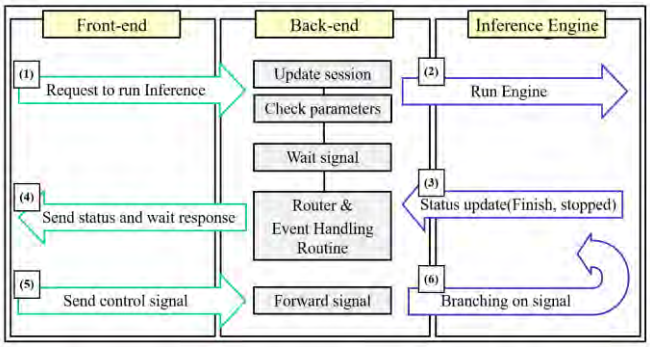
세션 관리자는 사용자의 상태를 세션을 이용하여 관리하며 사용자의 상태를 상태 변화 때마다 갱신한다. 세션은 프로세스 관리자가 프로세스의 정상 상태를 확인할 때도 사용된다. 이벤트 수신자가 처리하는 이벤트 루틴 중에서 상태 변화를 감지하면 세션 관리자가 세션의 상태를 업데이트한다.

E. 보조 기능

플랫폼에는 후위단에서 지원하는 보조 기능이 두가지 있다. 데이터 증가 기능은 학습을 위한 이미지를 변형시켜 학습 이미지의 개수를 증가시키며 공개 소스 라이브러리 Sharp[6]을 이용해 구현하였다. JSON 파서는 모델 파일과 가중치 파일을 Node.js 런타임에서 쉽게 사용하기 위하여 JSON 형식으로 변환시킨다. JSON 형식의 모델 파일은 신경망의 시각화와 추론 과정에서 신경망 층의 번호를 확인하고 각 층의 정보를 확인하기 위해 사용된다. 이진 형식의 가중치 파일은 차원을 가진 JSON 객체 형태로 변환되어 전위단과 후위단이 가중치 정보를 다룰 때 사용된다.

2.2 관리 모듈의 내부 동작

시뮬레이션 플랫폼에서 추론 기능이 동작하는 방식은 대표적인 관리 모듈로서 후위단의 역할을 보여준다. 그림 5 은 추론 엔진 구동 흐름을 보여준다. 전위단에서 추론 엔진을 구동하기 위한 메타데이터와 함께 보낸 추론 시작 요청(1)을 수신한 후위단은 추론 엔진을 자식 프로세스로 실행시킨다(2). 추론 엔진은 후위단의 내부 소켓 서버에 연결된다. 후위단은 세션을 참조하여 메타데이터를 세션에 갱신한다. 프로세스 관리자는 생성된 추론 프로세스가 종료될 때까지 주기적으로 추론 프로세스를 감시한다. 추론 프로세스가 실행될 때 사용하는 파라미터 파일들은 파일 관리자에 의해 독립적으로 관리된다. 따라서 여러 추론 프로세스들이 독립적으로 실행될 수 있다.



(그림 5) 추론 엔진 구동 흐름.

추론 엔진은 진행 상황에 따라 소켓 메시지를 후위단에게 보낸다(3). 후위단은 전위단에 프로세스의 상태를 웹 소켓을 통해 전송한다(4). 추론 엔진의 상태를 수신한 전위단에서 추론 프로세스를 제어하기 위한 신호를 후위단에 보내면(5), 후위단은 해당 사용자의 프로세스를 찾아 내부 소켓 포트를 통해 제어 메시지를 추론 프로세스에게 전달한다. 추론 프로세스는 전달받은 신호에 따라 추론의 계속 진행 또는 중지를 판단하여 분기한다(6). 이 때 프로세스 관리자는 프로세스 객체를 미리 저장해두고 사용한다. 추론 엔진의 구동이 후위단을 통해서 이루어지기 때문에, 후위단에서 추론 프로세스의 동작 범위를 제한함으로써 오동작을 제어할 수 있다.

3. 결론

부동소수점 연산기가 없는 실시간 임베디드 시스템 상에서의 추론은 신경망 모델의 양자화가 필수적이다. 본 논문은 신경망 모델의 양자화를 검증하는 시뮬레이터의 관리 모듈에 대해 기술한다. 관리 모듈은 Node.js 와 Express.js 엔진을 이용한 웹 서비스의 후위단으로 구현되었으며 웹 브라우저 상의 전위단과 서버의 추론 엔진을 연결하고 양자화 추론 동작을 관리한다.

관리 모듈은 크게 다섯개의 모듈로 구성된다: 라우터와 이벤트 수신자, 프로세스 관리자, 파일 관리자, 세션 관리자, 보조 기능. 관리 모듈의 동작으로 사용자는 웹 상에서 추론 엔진을 구동하며 추론 모델과 추론 과정의 데이터를 시각적으로 볼 수 있다. 프로세스과 파일 관리를 통하여 다중 사용자가 사용할 수 있는 환경이 구축되었다.

"본 연구는 과학기술정통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 연구결과로 수행되었음 "(2019-0-01816)

참고문헌

[1] TensorBoard: TensorFlow’s visualization toolkit [Internet], https://www.tensorflow.org/tensorboard

[2] Yeager, L., Bernauer, J., Gray, A., & Houston, M., “Digits: the deep learning gpu training system”, ICML 2015 AutoML Workshop, 2015.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", CVPR, 2016.

[4] Node.js [Internet], https://nodejs.org/en/.

[5] Express, Fast, unopinionated, minimalist web framework for Node.js [Internet], https://expressjs.com/.

[6] Sharp, High performance Node.js image processing [Internet], https://sharp.pixelplumbing.com/.