이기종 분산 처리 환경에서의 기기 자원 기반 추론모델 선정

박종빈*, 박효찬, 금승우
*한국전자기술연구원
*jpark@keti.re.kr

Resource aware Inference Model Selection for Heterogeneous Distributed Processing

Jongbin Park*, Hyochan Bak, Seungwoo Kum
*Korea Electronics Technology Institute

요 약

본 논문은 하드웨어 및 소프트웨어 구성 환경이 서로 다른 이기종 분산처리 환경에서 가용 기기 자원 정보를 토대로 기계학습 모델을 선택하고 제공하는 방법에 관한 것이다. 기기 자원 정보는 신경망 처리를 위한 범용 GPU의 장착 유무, CPU 아키텍처 및 모델, 메모리 용량, 네트워크 대역폭 정보를 사용한다. 본 논문에서는 제안 서비스의 가능성을 확인하기 위해 제어노드, 신경망 모델 관리 리포지토리, 추론을 위한 에지 노드들을 구성했고, 에지 노드에서 추론 속도 개선 효과를 측정했다.

Ⅰ. 서 론

이미지 추론을 위한 기계학습 모델은 특정 유형의 패턴을 인식하거나, 이미지에 포함된 객체를 유한개의 클래스로 분류한다 [1]. 딥러닝 기술의 등장 이후 이미지 추론을 위한 신경망 모델은 꾸준히 개선되고 있으며, 이미지 인식과 같은 일부 영역에서는 추론성능이 인간의 능력을 넘어선다 [2]. 그러나 여러 혼합기기, 특히 연산 능력이 제한되거나 네트워크 상황이 가변적인 모바일 및 에지 환경에서의 효과적인 추론모델의 선택 문제는 여전히 도전적이다. 본 논문에서는 이런 문제를 해결하기 위해 에지 기기의 연산 성능 및 정보를 이용하여 기기에 적합한 추론모델을 제공하여 처리하는 방법이 추론 속도의 개선 효과가 있었음을 제시한다.

Ⅱ. 본론

1. 전체 시스템 구성 및 성능 요약

에지 노드의 상태를 고려한 추론 모델 제공을 위한 시스템 환경은 그림 1과 같다. 제어노드, AI 모델 리포지토리, 데이터소스, 에지노드를 포함한다. 그림 2 좌측은 참조 시험환경(baseline), 우측은 제안 시험환경(advanced)을 보여준다. 모델 선택에서 차이를 갖는다. 모델 선택 시 제안 방법은 기기의 연산 자원 상태를 확인하여 모델을 결정한다. 그림 1의 AI 모델 리포지토리에는 Resnet [3], Mobilenet [4], Efficientnet [5] 구조로 사전에 학습한 이미지 추론모델을 탑재한다. 데이터소스에는 예측모델 및 분석 기술을 공유하고 경쟁하는 플랫폼인 kaggle에서 “imagenetmini-1000” 데이터셋의 validation 셋을 저장하고, 분석을 위해 색상정보 전처리를 수행한 후 에지 노드에 제공한다. 해당 데이터셋은 1,000개의 분류객체를 포함하며, 클래스당 약 3장씩의 이미지를 포함하여 총 3,923장으로 구성된다. 해당 구성에서 제안방법이 약 26%의 속도 개선을 보였다.

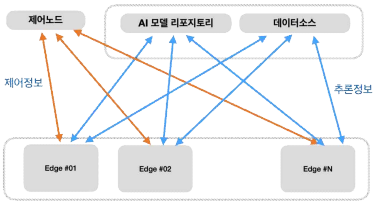

그림 1. 시스템 환경

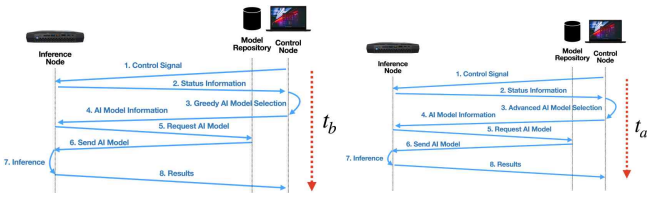

그림 2. 성능 비교를 위한 시험환경

Ⅲ. 결론

본 논문에서는 이기종 분산처리 환경에서 추론 모델을 제공함에 있어서 기기의 자원 상태정보를 활용하는 방법에 대해서 소개했다. 시험결과 종래의 단순 모델 제공 방법에 비해서 추론 성능의 큰 열화 없이 고속의 처리가 가능함을 확인했다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00907, 능동적 즉시 대응 및 빠른 학습이 가능한 적응형 경량 엣지 연동분석 기술개발).

참 고 문 헌

[1] Y. LeCun, Y. Bengio, J. Hinton, “Deep learning”, Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[2] W. Rawat, Z. Wang, “Deep convolutional neural networks for image classification: A comprehensive review”, Neural computation, vol. 29, no. 9, pp. 2352-2449, 2017.

[3] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[4] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications”, arXiv preprint arXiv:1704.04861, 2017.

[5] M. Tan, Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks”, International conference on machine learning, pp. 6105-6114, 2019