콘텐츠 인기도 예측 기반의 동적 메모리 할당 메커니즘

김가영 [1], 문정민 [1], 박나은 [2], 이일구 [3]
[1] 성신여자대학교 융합보안공학과 학부생
[2] 성신여자대학교 미래융합기술공학과 박사과정
[3] 성신여자대학교 융합보안공학과, 미래융합기술공학과 교수

20240925@sungshin.ac.kr, 20221093@sungshin.ac.kr, nepark.cse@gmail.com, iglee@sungshin.ac.kr

Content Popularity Prediction-Based Dynamic Memory Allocation Mechanism

Ga-Yeong Kim[1], Jung-Min Moon[1], Na-Eun Park[2], Il-Gu Lee[1,2]
[1] Dept. of Convergence Security Engineering, Sungshin Women's University
[2] Dept. of Future Convergence Technology Engineering, Sungshin Women's University

요 약

최근 산업에 활용되는 사물인터넷 장치의 수가 기하급수적으로 증가하면서 인공지능을 활용한 빅데이터 학습의 효율성과 성능이 중요해지고 있다. 그러나 기존의 메모리 할당 방식은 리소스 소모, 응답 시간 지연, 에너지 소모 측면을 해결하지 못하고 있다. 본 논문은 오토인코더와 LSTM 모델을 결합하여 데이터 요청 빈도 수에 따라 메모리를 동적 할당하는 CPP-DMA 모델을 제안한다. 실험 결과에 따르면 random method 대비 처리 시간을 30.3% 단축하고 메모리 사용량을 18.1% 절감할 수 있었다.

1. 서론

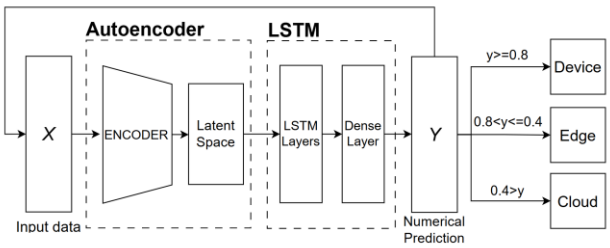
저전력 사물인터넷 장치가 산업에 널리 활용되면서 인공지능을 활용하여 빅데이터를 학습할 때 컴퓨팅 자원의 최적화 기술이 중요해지고 있다[1]. 그러나 종래의 정적 메모리 할당 기술은 리소스 사용과 에너지 소비 효율, 응답 지연 측면에서 비효율적이었다 [1, 2, 3].

최근 저전력 디바이스에서 인공지능을 활용하여 빅데이터를 학습 시 최적의 메모리 할당을 위해 다양한 연구가 진행되고 있다. 경매 기반 할당 연구 [2]는 엣지 노드가 실시간 수요에 따라 자원을 할당하지만, 과거 데이터 패턴에 의존하므로 실시간 적용이 어렵다. 5G 네트워크 자원 할당에 관한 연구 [3]에서는 GCN-LSTM(Graph Convolutional Network Long Short-Term Memory) 기반 예측을 활용하여 시스템 처리 효율을 향상시켰으나 사용자 접근성은 고려하지 못했다. 즉, 선행연구들은 과거 접근 패턴과, 실시간 데이터를 반영하지 못하고 복잡한 특성 분석으로 인한 지연 시간 문제를 해결하지 못했다[2, 3]. 이러한 종래 기술의 문제를 해결하기 위해 본 연구에서는 콘텐츠 인기도 예측 기반의 동적 메모리 할당 메커니즘 (Content Popularity Prediction-based Dynamic Memory Allocation, CPP-DMA)을 제안한다. 요청 빈도에 따라 콘텐츠 인기도를 예측하고 메모리를 동적으로 할당함으로써 에너지와 리소스를 효율적으로 관리할 수 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 제안하는 CPP-DMA 을 설명한다. 3 장에서는 제안하는 모델의 성능을 평가하기 위한 환경과 결과를 설명하고, 4 장에서는 결론을 맺는다.

2. 콘텐츠 인기도 예측 기반 동적 메모리 할당 메커니즘

본 절에서는 오토인코더와 LSTM 을 결합하여 실시간 메모리 할당을 수행하는 CPP-DMA 에 대해 설명한다. 제안 모델의 구조도는 그림 1 과 같다.



(그림 1) CPP-DMA 의 구조

입력 데이터는 오토인코더의 인코더를 통해 latent space 로 압축되며, 이 과정에서 데이터의 핵심 특징이 추출되고 노이즈가 제거된다. 축소된 특징은 LSTM 계층으로 전달되

어 데이터 접근 빈도를 학습한 후 Dense 레이어에 의해 0~1 범위의 정규화 된 예측 값을 출력한다. 예측 값이 0.8 보다 크면 디바이스 메모리, 0.4~0.8 범위이면 엣지 장치 메모리, 0.4 보다 작으면 클라우드 메모리에 할당한 후 데이터 변화에 따라 실시간으로 메모리를 할당한다.

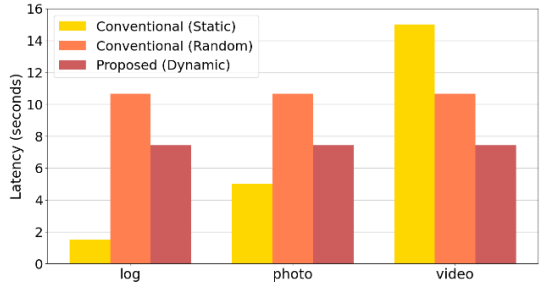
3. 성능 평가 및 분석

본 절에서는 제안 모델의 성능 검증을 위해 종래의 고정적으로 메모리 할당하는 static method 와 무작위로 메모리 할당하는 random method 를 설계하여 실행 시간, 메모리 사용량 측면을 비교했다. 각 모델은 파이썬으로 설계되었으며, 표 1 과 같은 실험 환경 조건을 두고 실험을 진행했다.

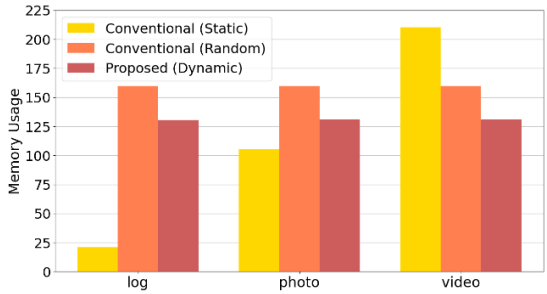
(표 1) 실험 환경 조건

Variable	Decryption
Request frequency range	Random (1-100s)
Memory size of data (MS)	Random (10-200MB)
Data type	Log/Photo/Video
Latency	Cloud (15s), Edge (5s), Device (1.5s)
Memory usage	Cloud (MS×2), Edge (MS×1), Device (MS×0.2)

Static method 는 콘텐츠 유형에 따라 고정된 위치에 저장하고, random method 는 클라우드:엣지:디바이스를 6:3:1 비율로 무작위로 할당했다. 실험은 메모리 할당에 따른 latency 와 memory usage 를 100 번 반복한 후의 평균값을 비교했다. 그림 2, 3 은 데이터 타입 별 latency 와 memory usage 를 비교한 결과이다.



(그림 2) 데이터 타입 별 latency 비교



(그림 3) 데이터 타입 별 memory usage 비교

그림 2 에서, random method 의 평균 처리 시간은 10.66 초였으나, CPP-DMA 는 7.42 초로 측정되었다. static method 는 콘텐츠 유형에 따른 성능 편차가 컸다. 반면, CPP-DMA 는 안정적인 처리 시간을 유지하며 일관된 성능을 제공했다.

그림 3 의 memory usage 비교에서는 CPP-DMA 가 random method 대비 약 30.3% 높은 성능을 보였다. 메모리 사용량 측면에서도 동일 모델 대비 약 18.1% 낮은 메모리 사용량을 기록했다. static method 는 콘텐츠 유형에 따른 큰 편차를 보였다. 즉, 실험을 통해 CPP-DMA 는 일관된 메모리 사용량을 유지하며 시스템 메모리 관리의 안정성과 예측 가능성을 향상시켰다.

4. 결론

인공지능을 활용한 빅데이터 학습의 효율성과 성능이 중요해지고 있다. 그러나 기존의 메모리 할당 방식은 데이터 접근성을 미반영과 지연 시간의 문제가 있었다. 본 논문에서는 효율적인 메모리 관리를 위해 CPP-DMA 를 제안했다. 실험 결과에 따르면 random method 대비 처리 시간을 30.3% 단축하고 메모리 사용량을 18.1% 절감시켰다. 후속 연구로는 제안 모델의 사용 빈도만을 기준으로 한 메모리 할당 방식을 개선하기 위해 수치형, 범주형, 추세 예측을 통한 동적 메모리 할당 메커니즘을 연구할 계획이다.

ACKNOWLEDGMENT

본 논문은 2024 년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재 4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Jeon, S.-E.; Oh, Y.-S.; Lee, Y.-J.; Lee, I.-G. Suboptimal Feature Selection Techniques for Effective Malicious Traffic Detection on Lightweight Devices. *Comput. Model. Eng. Sci.*, *140*(2), 1669-1687, 2024.

[2] Sushree S. S. Sudha, Aman Anand, Rishab Jain, Dipti Dash, Sujata Swain, Anjan Bandyopadhyay. Auction Based Dynamic Resource Allocation in Edge Computing. *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, Bhubaneswar, India, 2024, 6.

[3] Xu Gao, Jianfeng Wang, Mingzheong Zhou. The Research of Resource Allocation Method Based on GCN-LSTM in 5G Network. *IEEE Communications Letters*, *27*(3), 926–930, 2023.