IoT 디바이스의 딥러닝 기반 데이터 전처리를 이용한 효율적인 오디오 AI를 위한 엣지 컴퓨팅 모델

Edge Computing Model for Efficient Audio AI Using Deep Learning Based Preprocessing on IoT Devices

요 약

최근 IoT 장치의 성능이 향상됨에 따라 엣지 IoT 환경에서 다양한 종류의 데이터를 대량으로 수집할 수 있게 되었다. 특히, IoT 장치들의 다양한 센싱 정보 중에서 오디오 데이터를 활용하여 생태계 분석, 감정 분석, 상황 분석 등을 포함한 지능형 모니터링을 기능을 구현하는 연구가 활발히 진행되고 있다. 그러나, 대부분의 연구에서는 소리 인식 및 잡음 제거와 같은 데이터 전처리 과정이 일반적으로 엣지 서버에서 수행되고 있는데, 이는 엣지 서버의 과부하와 및 네트워크 트래픽 증가와 같은 문제점을 동반한다. 본 논문에서는 이를 해결하면서 딥러닝 기반 음성 품질 향상을 효율적으로 수행할 수 있는, IoT 기기의 오디오 전처리를 이용한 엣지 컴퓨팅 기법을 제안한다. 제안 기법에서 IoT 장치는 (엣지 서버에 오디오 데이터를 전달하기 전에) 미리 사전에 학습된 CNN 기반 소리 인식 모델을 이용하여 수집 대상 소리를 판별하고, AECNN 기반 잡음 감소 모델을 사용하여 소리 데이터의 잡음을 감소시킨 뒤, 이렇게 전처리 된 데이터를 엣지 서버로 전송한다. 이를 위해, 즉 상대적으로 성능이 낮은 IoT 기기에서 CNN과 AECNN 모델을 적용하기 위해, 기존 기법보다 경량화된 딥러닝 모델을 사용하였다. 시뮬레이션 결과 제안된 엣지 컴퓨팅 기법이 기존 기법보다 네트워크 통신량, 엣지 서버 사용률 측면에서 현저히 감소함을 확인하였다.

1. 서 론

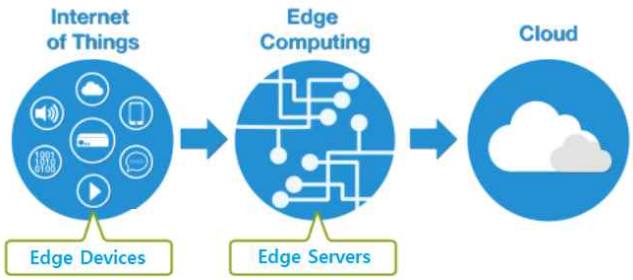

그림 1. 엣지 컴퓨팅 모델

엣지 컴퓨팅은 IoT 기기 근거리에 있는 가까운 네트워크 가장자리(엣지)에서 데이터를 실시간 처리하는 기술이다. 사물인터넷 기기나 센서가 모두 클라우드에 직접적으로 연결되는 것을 막기 위해 등장했다. 이 기술에서 엣지 디바이스들은 효율적으로 데이터를 수집하고 이를 엣지 서버로 전송하는 역할을 수행한다[1]. 엣지 서버에서는 수집된 데이터를 처리하기 위해서 정형화되지 않은 데이터를 전처리하는 과정을 포함하여 응용을 위한 다양한 AI 학습을 진행한다. 그런데, 이 과정을 엣지 서버에서 온전히 수행할 경우 엣지 서버로의 네트워크 트래픽 부하 및 시스템 컴퓨팅 부하가 발생하여 엣지 서버에서 처리해야 할 많은 작업들을 수행하는데 어려움이 있을 수 있다. 이러한 문제점을 해결하기 위해 여러 연구가 활발히 수행되었고, 특히 하드웨어의 성능이 발전함에 따라 데이터 전처리 과정을 엣지 디바이스에서 일부 수행하는 기법들이 제안되고 있다[2].

한편, 소리 데이터는 다양한 기계 학습 응용을 위한 데이터셋으로 유용하게 사용되고 있다. 잡음 감지 시스템을 위한 도시 소리 분류, 새의 생태계 분석을 위한 환경 소리 분류, 어조를 통해서 사람의 감정을 분석하는 감정 분석 등 많은 분야에서 활용되고 있다[3][4][5]. 다양한 종류의 소리 데이터를 기계 학습에 사용하기 위해서는 양질의 소리 데이터를 분리하여 확보하는 기술[6][7]이 필요한데, 이는 수집 대상 소리 이외의 데이터나 잡음이 포함될 경우, 학습의 정확도가 떨어질 수 있기 때문이다. 따라서 원하는 소리 이외의 잡음을 제거하여 명료한 오디오 데이터를 이용하여 학습의 정확도를 높이는 기술이 필요하다.

본 논문에서는 위의 두 가지 기술 동향을 바탕으로, 엣지 서버에서의 음성 품질 향상 딥러닝 모델을 위하여, 엣지 IoT 디바이스에서 미리 양질의 소리 데이터 추출을 위한 데이터 전처리를 수행하는 엣지 컴퓨팅 모델을 제안한다. 엣지 디바이스에서 데이터 전처리를 수행함으로써, 엣지 서버의 컴퓨팅 부하를 줄이고, 또한 전처리 된 데이터만을 엣지 서버로 전송함으로써 엣지 서버로의 네트워크 부하를 줄 일 수 있다. 제안하는 모델에서는, 일반적으로 서버에서 수행하는 전처리 과정을 엣지 디바이스에서 수행하기 위하여, 더욱 경량화된 딥러닝 모델을 사용한다. 구체적으로 엣지 디바이스의 전처리 과정은

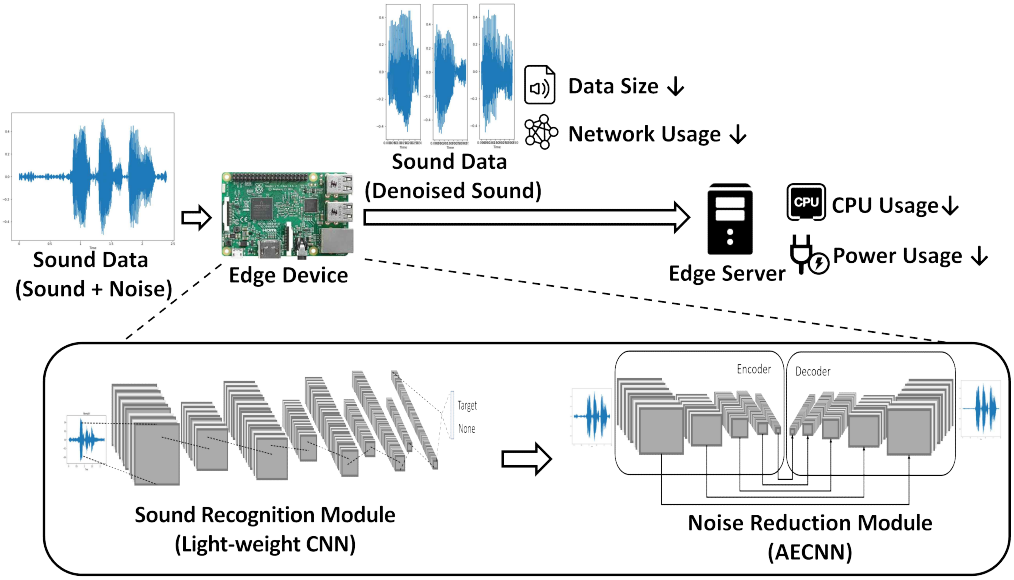

그림 2. 제안 기법의 개요

경량화 CNN(Convolutional Neural Networks) 기반 소리 인식 모델을 활용해 원하는 소리를 판별하고, AECNN(Auto- Encoder Convolutional Neural Networks) 기반 잡음 감소 모델[8]을 활용하여 명료한 소리 데이터를 얻는다.

본 논문의 순서는 다음과 같다. 2장에서 제안 기법에 대해 기술하고, 3장에서 실험 환경과 성능 분석에 대해서 기술한다. 4장에서 결론 및 향후 연구 과제로 마무리한다.

2. 제안 기법

그림 1은 제안된 엣지 컴퓨팅 모델의 전체적인 개요이다. IoT 엣지 디바이스가 소리 데이터를 수집하면 엣지 서버로 데이터를 전송하기 전에 전처리 과정으로 소리 인식 모듈(Sound Recognition)과 잡음 제거 모듈(Noise Reduction)을 거친다. 각 모듈은 수집 대상 소리 판단과 잡음을 제거하는 역할을 한다. 각 모듈을 거치고 나면 명료한 소리 데이터를 얻을 수 있고, 이 전처리 된 데이터를 엣지 서버로 전송한다.

2.1. 소리 인식 모듈

소리 인식 모듈은 수집된 데이터가 수집 대상 소리인지 판단하는 역할을 한다. 인식을 위한 모델은 CNN 기반 딥러닝 모델을 사용하며[9], 엣지 디바이스에서 구동될 수 있게 CNN 구조를 경량화하여 설계하였다. 구체적인 CNN의 구조는 3x3의 필터를 갖는 5개의 합성곱 층과 2x2의 필터를 갖는 4개의 최대 풀링 층으로 구성된다. 또한, 활성화 함수는 ReLU를 사용하였으며 입력의 크기는 깊이가 1인 98x128 멜-스펙트로그램이다. 이후 FC(Fully Connected) 층을 softmax 함수와 함께 이용하여 예측값을 출력한다[6]. CNN 모델의 학습은 많은 GPU 연산을 요구하기 때문에 별도의 환경에서 사전 학습한다.

전술한 바와 같이 사전 학습된 CNN 기반 모델을 사용하여 수집된 데이터가 수집 대상 소리인지 판단하며, 수집 대상 소리라고 판단된 데이터는 다음 단계인 잡음 제거 모듈로 전달하고, 그렇지 않은 데이터는 제거한다. 이로 인해 불필요한 프로세싱과 데이터의 전송은 하지 않는다.

2.2. 잡음 감소 모듈

잡음 감소 모듈은 소리 데이터의 잡음을 감소시켜 명료한 소리 데이터를 만드는 역할을 한다. 잡음 감소를 위한 모델은 AECNN 기반 딥러닝 모델을 사용하며, 엣지 디바이스에서 구동될 수 있게 AECNN 구조를 경량화하여 설계하였다. AECNN은 CNN에서 인코더 압축을 통해 하위 수준의 세부 정보 손실을 방지하기 위해 각 인코딩 계층을 해당 디코딩 계층에 연결하는 잔차 연결을 사용한다[7]. 구체적인 AECNN의 구조는 다음과 같다. 실시간 잡음 감소를 위해 지연 시간은 8ms로 설정하였으며, 256 프레임의 입출력 데이터 크기와 15x15 필터를 갖는 5개의 인코더 층과 5개의 디코더 층을 사용하고, 활성화 함수로 PReLU를 사용한 구조에서 총 257,928개의 파라미터를 사용한다. AECNN 모델의 학습도 마찬가지로 많은 GPU 연산을 요구하기 때문에 별도의 환경에서 진행한다.

전술한 바와 같이 사전 학습된 AECNN 기반 잡음 감소 모델을 사용하여 소리 인식 모듈에서 전달받은 데이터를 처리한다. 이 과정을 통해 명료한 소리 데이터를 얻을 수 있고, 이 소리 데이터를 엣지 서버로 전송하게 된다. 이로 인해 엣지 서버에서는 잡음 감소와 같은 전처리 과정을 수행하지 않아도 되기 때문에 서버의 부하가 줄어들게 되고 전송되는 데이터 트래픽의 양도 줄어들게 된다.

3. 실험 결과

3.1. 실험 환경

본 실험에서 엣지 디바이스는 라즈베리 파이 4를 사용하였고, 엣지 서버와 엣지 디바이스 간의 통신은 블루투스를 사용하였다. 또한 CNN, AECNN 모델 학습을 위한 데이터세트로 [10]의 데이터세트를 사용하였다. 이 데이터세트는 음성 인식 및 잡음 감소 학습을 위한 데이터세트로 잡음이 포함된 음성과 잡음이 제거된 음성으로 이루어져 있다.

3.2. 실험 내용

3.2.1. 모델 학습

모델 학습은 별도의 서버에서 진행한다. 데이터세트를 16kHz로 샘플링하여 음성 데이터를 1초 단위로 분할한 후 학습을 진행한다. 소리 인식 CNN 모델은 2.1.의 구조에 따른 잡음이 포함된 음성을 사용하고, 배치 크기 16, 에폭 40으로 설정하여 학습을 진행한다. 잡음 감소 AECNN 모델은 잡음이 포함된 음성과 잡음이 포함되지 않은 음성을 학습에 사용하고 2.2.의 구조에 배치 크기 100, 에폭 40으로 설정하여 학습을 진행한다.

3.2.2. 실험 과정

데이터세트에서 392개의 총 1,152초 길이의 음성 데이터와 잡음만 있는 2,448초의 데이터를 결합하여 1시간 길이의 테스트 데이터를 생성한다. 테스트 데이터를 1초 단위로 분할하고 CNN 모델을 사용하여 샘플링된 데이터가 수집 대상 음성인지 판단하고 수집 대상 음성이 아닌 경우 삭제한다. 이어 연속된 오디오 파일을 결합한 후 AECNN 모델을 사용하여 잡음 감소를 수행한다.

3.3. 성능 평가

엣지 서버에서 수행했을 경우와 엣지 디바이스에서 수행했을 경우의 엣지 서버와 엣지 디바이스에서 증가한 평균 CPU 점유율과 엣지 서버가 수신하는 데이터 크기를 비교하였다. 결과는 표 1과 같다.

표 1. 전처리 위치에 따른 성능 지표

전처리 위치 / 성능 지표	엣지 서버	엣지 디바이스
엣지 서버 평균 CPU 점유율	115.17%	13.98%
엣지 서버 수신 데이터 크기	110MB	31MB

엣지 서버의 평균 CPU 점유율은 엣지 서버에서 전처리할 경우 115.17%, 엣지 디바이스에서 전처리할 경우 13.98%로 후자의 경우 엣지 서버의 평균 CPU 점유율이 101.19% 감소함을 확인할 수 있다. 엣지 서버가 수신하는 데이터 크기는 엣지 서버에서 전처리할 경우 110MB, 엣지 디바이스에서 전처리할 경우 31MB로 후자의 경우 데이터 크기가 79MB 감소함을 확인할 수 있다. 따라서, 엣지 디바이스를 통해 전처리 과정을 거치는 것이 엣지 서버의 사용량 및 네트워크 사용량을 감소시켜 엣지 서버의 부하를 감소시키는 것을 확인할 수 있었다.

4. 결론 및 향후 연구 과제

본 논문에서는 소리 데이터를 수집하는 엣지 IoT 환경에서 엣지 서버의 사용량 및 네트워크 사용량을 감소시키기 위해 엣지 디바이스에서 수집한 소리 데이터에 대한 소리 인식 및 잡음 감소를 수행하는 기법을 제안한다. 엣지 서버에서 수행할 전처리 작업을 엣지 디바이스에서 수행한다면 엣지 서버가 전처리에 수행할 자원을 다른 작업에 할당할 수 있어 더 효율적인 IoT 환경을 운영할 수 있을 것으로 기대된다.

다만, 엣지 디바이스에서 소리 인식, 잡음 감소를 수행할 경우 엣지 디바이스의 사용량이 커짐에 따라 에너지 문제가 발생할 수 있다. 따라서, 에너지 수집형 엣지 디바이스를 사용하여 에너지가 충분할 경우 전처리 작업을 충분히 수행하고 엣지 디바이스의 에너지가 충분하지 않을 경우 전처리 과정을 줄인 채로 엣지 서버로 전송하는 방식에 대한 연구가 가능할 것이라 생각한다.

참 고 문 헌

[1] C. Savaglio, et al, “Data mining at the IoT edge”, 2019 28th International Conference on Computer Communication and Networks(ICCCN), pp. 1-6, 2019.

[2] F. Lin, Y. Zhou, X. An, I. You, and K. K. R. Choo, “Fair resource allocation in an intrusion-detection system for edge computing: Ensuring the security of Internet of Things devices”, IEEE Consumer Electronics Magazine, vol. 7, no. 6, pp. 45-50, 2018.

[3] Sonmez, Y. Ülgen, and A. Varol, “New trends in speech emotion recognition”, 2019 7th International Symposium on Digital Forensics and Security(ISDFS), pp. 1-7, 2019.

[4] Farnsworth, Andrew, et al,. “Automating Acoustic Monitoring of Nocturnally Migrating Birds: BirdVox and the Integration of Citizen Science and Radar Data to Enhance Evolving Paradigms”, AGU Fall Meeting Abstracts, vol. 2019, pp.B13A-06, 2019.

[5] Bello, P. Juan, et al, “Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution”, Communications of the ACM, vol. 62, no. 2, pp. 68-77, 2019.

[6] Yuh A. H., and Kang S. J., “Real-Time Sound Event Classification for Human Activity of Daily Living using Deep Neural Network”, 2021 IEEE International Conferences on Internet of Things(iThings) and IEEE Green Computing & Communications(GreenCom) and IEEE Cyber, Physical & Social Computing(CPSCom) and IEEE Smart Data(SmartData) and IEEE Congress on Cybermatics (Cybermatics), pp. 83-88, 2021.

[7] Turpault, Nicolas, et al. “Improving sound event detection in domestic environments using sound separation”, arXiv preprint arXiv:2007.03932, 2020.

[8] Drakopoulos, Fotios, D. Baby, and S. Verhulst, “Real-time audio processing on a Raspberry Pi using deep neural networks”, 23rd International Congress on Acoustics (ICA 2019), pp. 2827-2834, 2019.

[9] A. Saad, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network”, 2017 international conference on engineering and technology(ICET), pp. 1-6, 2017.

[10] Valentini et al, “https://datashare.ed.ac.uk/handle/10283/1942”, 2016.