초소형 머신 러닝 기반 신호 검파기 모델

김연군, 이정우
서울대학교 전기정보공학부
ygoonkim@cml.snu.ac.kr, junglee@snu.ac.kr

Tiny Machine Learning based Symbol Detector

Yeongoon Kim, Jungwoo Lee
Department of Electrical and Computer Engineering, Seoul National University

요 약

ViterbiNet, BCJRNet 등의 딥 러닝 기반 채널 상태 정보(Channel State Information) 예측 및 신호 검파기 기술은 Viterbi 알고리즘 등의 기존 알고리즘에 필적하는 성능을 내고 있지만, 이를 실제 하드웨어에 구현하기 위한 연구는 여전히 추가 진척이 필요하다. 딥 러닝 측면에서 이러한 연구들은 초소형 머신 러닝으로 분류되어 연구되고 있는데, 본 논문에서는 기존 초소형 머신 러닝에서 연구된 모델들을 ViterbiNet에 적용하였을 때, 모델의 크기를 1/3, 이론적으로는 1/6으로 줄이고도 학습 시간 및 성능 면에서 유의미한 상승이 이뤄짐을 발견했다.

Ⅰ. 서론

최근 대규모 언어 모델 (Large Language Model)의 도래를 필두로 다양한 분야에서 머신 러닝 기술을 적용하려는 연구가 진행되고 있다. 그 중 디지털 통신 분야에서는, 실제 디지털 신호의 채널을 통한 송수신 통신 환경에서 필수적인 계산이나 발생할 수 있는 다양한 불안정성을 추정 및 수정하는 알고리즘을, 딥 러닝 기반 모델로 대체하여 추정 정확도를 높이려는 연구들이 주로 이루어지고 있다. 그 중 대표적으로 연구되고 있는 기술은 신호 검파기(Symbol Detector) 기술로, 노이즈가 포함된 채널 송수신 환경에서 에러가 추가된 입력으로부터 출력단에서 에러를 추정 및 제거하는 기술을 의미한다. 기존에는 이를 해결하기 위해 Viterbi[1], BCJR[2] 알고리즘 등의 model-based 추정 알고리즘 등을 사용해 왔지만, 최근 해당 알고리즘에서 채널 상태 정보를 추정하기 위해 ViterbiNet[3], BCJRNet[4] 등의 딥 러닝 기반 신호 검파기 기술이 제안되고 있다.

이러한 딥 러닝 기반 기술을 반도체, IoT 디바이스 또는 네트워크 칩에 적용하기 위해서는 인공신경망 경량화 기술의 적용이 필수적인데, 이러한 방법론들은 주로 초소형 머신 러닝(Tiny Machine Learning)이라는 주제하에 연구되고 있다. 현재는 모듈 외부에서 가지치기(Pruning), 신경망 아키텍쳐 탐색(Neural Architecture Search), 신경망 양자화(Network Quantization), 지식 증류(Knowledge Distillation) 등의 기법을 적용하여 모델의 크기를 최소한으로 줄이고, 이를 아두이노 등의 소형 칩에서 추론 가능하도록 하는 연구가 주를 이루고 있으며, 추가로 소형 칩에서의 학습을 위한 분산 컴퓨팅(Distributed Computing), 양자 컴퓨팅(Quantum Computing) 및 학습 (Quantized Training), 희소 학습(Sparse Training) 등의 기법이 연구되고 있다.

이와 같은 딥 러닝 기반 신호 검파기 기술의 실적용 가능성을 알아보기 위해 본 연구에서는 구현한 Viterbinet Pytorch 시뮬레이터에 기반하여, TinyML 적용에 필요한 다양한 기법들, 예를 들어 양자화 인지 학습(Quantization Aware Training), 가지치기 기법들의 복합적 적용, 기존 경량화 RNN Cell 기법 적용 등을 통해 성능 및 모델 크기, 추론 속도 등을 확인하여 딥 러닝 기반 신호 검파기 기술의 H/W 구현 가능성을 검증한다.

Ⅱ. 본론

딥 러닝 기반 신호 검파기 기술 Symbol Detection 알고리즘 중에서 가장 잘 알려진 기술인 Viterbi 알고리즘은 최대 우도 추정(Maximum Likelihood, ML)과 최대 사후 확률 추정(Maximum a Posterior, MAP)에 기반하여, 출력 코드와 CSI로부터 계산된 finite-memory channel의 path cost 또는 factor graph가 주어질 때 가장 error rate가 적은 예상 입력 신호를 추정하는 알고리즘이다. 이 때 CSI가 요구되는 path cost 계산 및 factor graph 구축을 입출력 코드로 구성된 신호 데이터로 학습 가능한 DNN(Deep Neural Network)로 대체하여, CSI가 주어지지 않고서도 Viterbi 알고리즘을 적용할 수 있도록 하는 것이 ViterbiNet의 핵심 아이디어이다.

딥 러닝 네트워크 경량화 대표적인 네트워크 경량화 기술에는 상대적으로 덜 중요하다고 여겨지는 네트워크 셀이나 연산 과정을 랜덤하게, 또는 신경망 아키텍쳐 탐색 과정을 통해 찾아내 제거하여 네트워크의 경량화를 달성하는 가지치기 기법, 그리고 네트워크의 소수 자료형 파라미터를 정수 자료형 파라미터로 양자화하는 기술인 네트워크 양자화 기법이 있다. 하지만 초소형 머신 러닝 환경에서는 단순히 위 기법들을 범용적으로 적용하는 것에는 한계가 있으며, 사용하려는 네트워크의 종류 및 하드웨어 사양에 딸느 세심한 조정이 필수적이다.

초소형 머신 러닝 초소형 머신 러닝의 대표적인 연구로는 MobileNet[5], EfficientNet[6], MCUNet[7] 등이 있다. 이러한 연구들은 주로 CNN 기반 이미지 분류 task를 대상으로, covolution 등 사용되는 연산 과정에 다양한 기법을 적용하여 추론 시간 최소화 및 1MB 이하의 메모리 기기에 동작이 가능하도록 하는 것을 목표로 한다. 하지만 ViterbiNet에는 LSTM

기반 네트워크가 들어가기 때문에, 직접적인 적용이 불가능하다. 이에 대한 대안으로 FastRNN, FastGRNN[8] 등의 경량 RNN 모델을 활용할 수 있다.

실험 방법 본 실험에서는 기존에 구현된 Matlab 기반 ViterbiNet을 기초로 하여 Pytorch 기반 시뮬레이터를 구현하였다. 이후 CSI 예측에 사용되는 ViterbiNet 내부의 LSTM 네트워크를. 경량 모델인 FastRNN, FastGRNN으로 대체하였을 때의 성능, 추론 시간, 모델 크기 등을 측정하여 성능을 비교하였다. 또한 가지치기 기법을 적용하였을 때, 가지치기 비율과 성능 하락 정도를 측정하여 최적의 가지치기 비율을 찾아보았다. 실험 결과는 다음과 같다.

모델	LSTM (Baseline)	FastRNN	FastGRNN
모델 크기(KB)	308	107	106
학습 시간(s)	15.66	5.12	7.8
추론 시간(s)	~4s	~4s	~4s

표 1. 모델 종류에 따른 Viterbinet Specification

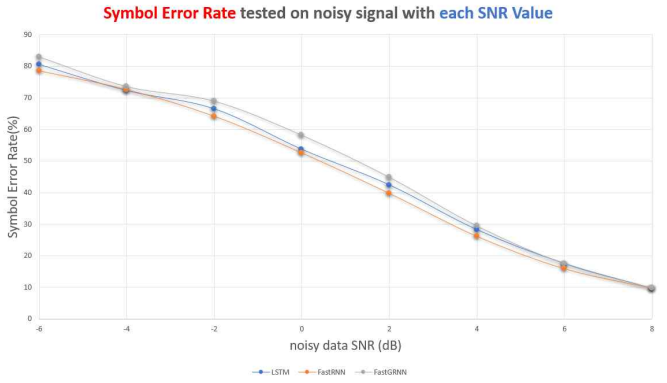

차트 1. 각 SNR value를 가지는 noisy data에 대한 Symbol Detecting 시 SER (파랑 - LSTM. 주황 - FastRNN, 회색 - FastGRNN)

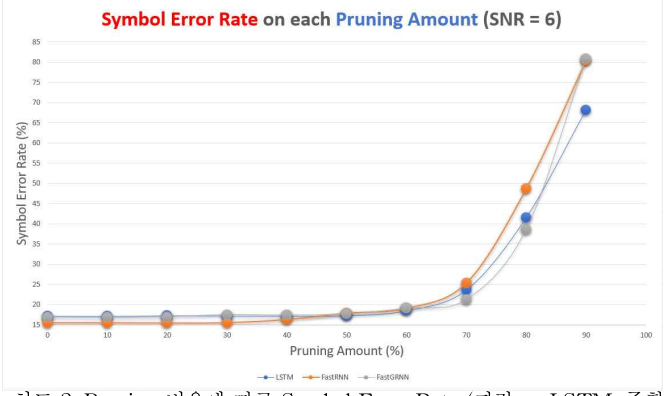

차트 2. Pruning 비율에 따른 Symbol Error Rate (파랑 - LSTM. 주황 - FastRNN, 회색 - FastGRNN)

실험결과 및 논의 먼저 모델 간 비교 실험 결과, 기존 ViterbiNet 내 LSTM을 FastRNN으로 대체하였을 때 모델 크기는 약 1/3으로 감소하였고, 그에 비례하여 학습 시간 또한 1/3으로 감소하였다. 또한 동시에 Symbol Error Rate 또한 모든 SNR 환경에서 2~3% 가량 감소하여, 더 작은 모델을 적용하고도 더 높은 성능을 보여주었다. 또한 해당 FastRNN 대체 모델에서 50%의 파라미터를 가지치기 하여도 SER이 0.4%, 60% 가지치기 환경에서 0.7% 정도로 미약하게 상승하는 것을 관측하였고, 따라서 시뮬레이터 상으로 40~50KB 내외의 인공신경망 네트워크만을 가지고도 ViterbiNet 등의 딥 러닝 기반 신호 검파기 기술을 구현할 수 있다는 것을 보였다.

III. 결론

본 논문에서는 초소형 머신 러닝 분야에서 연구된 FastRNN, FastGRNN을 딥 러닝 기반 채널 상태 정보 예측 및 신호 검파기 기술인 ViterbiNet에 적용하여, 모델의 크기를 1/3~1/6으로 줄이고도 ViterbiNet이 정상적으로 동작함을 확인하였고, 이에 따라 100KB 이내의 소형 메모리를 가진 네트워크 칩에도 딥 러닝 기반 신호 검파기 기술을 사용할 수 있음을 보였다.

ACKNOWLEDGMENT

This work is in part supported by Samsung Electronics Co.,Ltd(Contract ID: MEM210728_0001), National R&D Program through the National Research Foundation of Korea (NRF, 2021M3F3A2A02037893, 2021R1A2C2014504), Institute ofInformation & communications Technology Planning & Evaluation (IITP, 2021-0-02068) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21 FOUR program.

참 고 문 헌

[1] A. Viterbi. “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. IEEE Transactions on Information Theory, vol. 13, no. 2, pp. 260-269. 1967.

[2] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, “Optimal decoding of linear codes for minimizing symbol error rate,” IEEE Transactions on Information Theory, vol. 20, no. 2, pp. 284-287, 1974

[3] N. Shlezinger *et al.* "ViterbiNet: Symbol Detection Using a Deep Learning Based Viterbi Algorithm," 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1-5, 2019.

[4] N. Farsad *et al.* "Data-Driven Symbol Detection Via Model-Based Machine Learning," 2021 IEEE Statistical Signal Processing Workshop (SSP), pp. 571-575, 2021.

[5] A. Howard *et al.* “Searching for MobileNetV3,” Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), 2019.

[6] M, Tan and Quo V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” Proceedings of the 36th International Conference on Machine Learning(ICML), 2019.

[7] J. Lin *et al.* “MCUNet: Tiny Deep Learning on IoT Devices,” Advances in Neural Information Processing Systems (NeurIPS), 2020.

[8] K. Aditya, *et al.* "FastGRNN: A Fast, Accurate, Stable and Tiny Kilobyte Sized Gated Recurrent Neural Network“ Advances in Neural Information Processing Systems(NeurIPS) 31, 2018.