신경망 기반 복호기의 매개변수 양자화에 따른 성능 분석

나혜연, 유민희, *박호성
전남대학교 컴퓨터정보통신공학과
201660@jnu.ac.kr, 201550@jnu.ac.kr, hpark1@jnu.ac.kr

Performance Analysis on Parameter Quantization in Neural Decoder

Hyeyeon Na, Minhui Yu, *Hosung Park
Department of Computer Engineering
Chonnam National University

요 약

본 논문에서는 딥러닝 기반의 복호기에서 양자화 레벨에 따른 매개변수 양자화가 오류율에 미치는 영향을 연구한다. 제시된 딥러닝 기반의 신경망 구조는 각 반복의 노드마다 가중치와 편향치를 할당한 형태이다. 복호기에서 가중치와 편향치를 모두 양자화하지 않은 결과와 모두 양자화를 적용한 결과를 비교한다. 이 연구에서 양자화는 1~6 비트 범위의 결과를 비교하였고, 복호기는 비트 오류율(bit error rate: BER)과 블록 오류율(block error rate: BLER)을 측정함으로써 성능을 판단한다.

Ⅰ. 서론

최근 5G NR(New Radio) 시스템에서 데이터 채널의 부호화 방식으로 저밀도 패리티 체크(Low Density Parity Check: LDPC) 부호가 선택되고 있다. 또한 딥러닝 분야에서 모델을 압축하는 네트워크 경량화 기술에 대한 관심이 높아지고 있다. LDPC 부호의 복호기 알고리즘 중에서 Neural MS(Min-Sum) 디코딩은 각 레이어의 노드별로 가중치와 편향치를 할당하여 학습하는 방식을 가진다. 학습에 사용되는 매개변수인 가중치와 편향치의 경량화는 모델 압축의 방법이 될 수 있다.

대표적인 네트워크 경량화 기술로 양자화 기법이 있다. 기존 양자화 기법은 실수형 변수를 정수형 변수로 변환하는 기술이다. 본 연구에서는 실수형에서 정수형으로의 변환보다 실수형에서 실수형으로의 변환이 모델에 적합하므로 이와 같이 적용한다. 딥러닝 모델에서 학습에 사용하는 가중치와 편향치를 양자화 기법에 적용함으로써 모델 크기 최소화와 메모리 최적화를 달성할 수 있다. 이를 토대로 본 논문은 Neural MS 디코딩에서 양자화 레벨을 1~6 비트로 변환함에 따라 복호기의 성능에 어떤 영향을 미치는지를 연구한다.

Ⅱ. 본론

1. 매개변수 학습

본 연구는 그림 1과 같은 신경망 네트워크 구조를 사용하며, 이는 프로토그래프 기반 LDPC 부호에서 효율적으로 동작할 수 있도록 제안되었다[1]. 학습 과정에서의 기울기 소실 문제를 해결하고 다양한 부호길이와 부호율에 대한 일반화를 위하여 두 가지의 학습 방법을 사용한다.

첫 번째로 부호길이/부호율 호환 학습이다. 리프팅 과정에서 매개변수 공유 알고리즘을 통해 하나의 매개변수 배열은 동일한 기본 부호어로부터 파생된 여러 개의 리프팅된 부호어에 적용될 수 있다.

두 번째는 반복별 탐욕적인 학습 방법이다. 일반적으로 딥러닝 신경망에서는 기울기 소실 문제를 해결하기 위하여 다중 손실 함수를 사용하는데, 우리는 반복별 탐욕적인 학습 방법을 사용한다. 한번 레이어가 학습되면 해당하는 가중치와 편향치가 다음 반복을 위해 고정되며, 매번 마지막 레이어의 가중치와 편향치만 학습 가능하다.

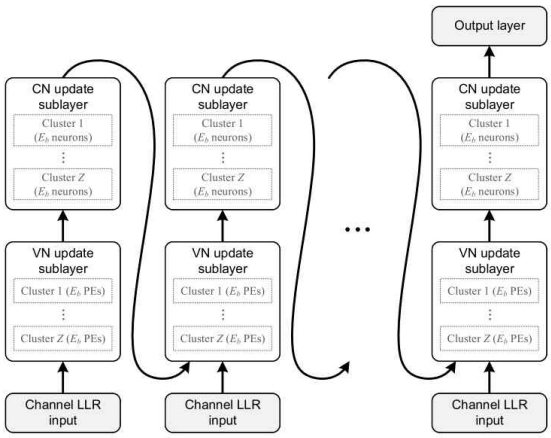

그림 1 . Neural MS 네트워크의 구조 [1]

2. 매개변수 양자화

본 연구에서 양자화 방식으로 선택한 PTQ(Post Training Quantization)는 학습이 끝난 매개변수에 대해 양자화를 적용하는 기법이다. 본 연구에서는 양자화 이전의 기존 매개변수 범위를 고려하여 실수형으로 양자화 구간을 정의한다. 각 구간은 양자화 레벨의 비트 수에 따라 1~6 비트를 사용하며, 이로 인해 매개변수의 값을 2^1~2^6개의 구간으로 분류했다. 이 과정에서 균일 양자화 기법을 사용하여 표본 값이 속한 모든 양자화 레벨 구간을 균일한 간격으로 분배하였다. 각 구간의 평균을 대푯값으로 선택하고, 각 매개변수는 해당하는 구간의 대푯값으로 변환된다.

Ⅲ. 모의실험

1. 실험환경

본 연구는 복호 시에 수많은 복호어 대신 0으로 구성된 단 하나의 복호어를 사용하였다.

학습률	0.001
리프팅 인자(Z)	16
학습 반복 횟수	25
복호 반복 횟수	25
(n,k)	(52,42)

표 1. 실험 환경

본 연구는 5G NR의 기본 그래프2에 해당하는 프로토 그래프 기반 LDPC 부호를 이용하여 모의실험을 진행하였다. 표 1에서와 같은 값을 사용하여 매개변수를 학습하였으며 활성화 함수로는 ReLU를 사용하였다. 실험은 가산성 백색 가우시안 잡음(Addictive White Gaussian Noise: AWGN) 채널에서 PBRL LDPC 부호어를 사용하여 진행하였다. 25번의 반복에 해당하는 모든 레이어들의 노드에 각각 다른 가중치와 편향치를 할당하였고, 매개변수 공유 알고리즘과 반복별 탐욕적인 학습 방법을 사용하여 최적의 값으로 학습하였다. 학습에는 GPU를 사용하였다. 학습을 완료한 후 양자화 레벨 2, 4, 8, 16, 32, 64에 맞춰 구간을 정의하였고 각 구간에 해당하는 대푯값으로 양자화를 하였다.

2. 실험

본 실험은 양자화 레벨에 따른 매개변수 양자화를 성능 측정함으로써 시뮬레이션 결과를 나타낸다. 양자화 레벨은 1~6 비트를 기준으로 2^1~2^6 레벨을 의미한다. 이를 통해 각 레벨의 양자화 결과가 복호기에 미치는 영향을 도표로 표현하여 최종적인 성능을 비교 분석한다.

양자화 레벨에 따른 비트 오류율 결과와 블록 오류율 결과는 각각 그림 2, 그림 3에 제시되어 있다. Neural MS 복호기는 무작위로 샘플을 선택하여 시뮬레이션을 진행한다. 성능을 비교하기 위해 비트 수를 조절해가며 학습에 사용한 매개변수를 양자화하였다.

그림 2와 3을 참고하여 성능 차이를 시각적으로 확인할 수 있다. 양자화 레벨이 2일 때, 다른 레벨에 비해 성능이 현저히 낮은 것을 알 수 있다. 이 경우에는 양자화 레벨을 극단적으로 제한하였으므로 이러한 결과를 보이는 것이다. 양자화 레벨이 2인 경우를 제외한 나머지 레벨(4, 8, 16, 32, 64)들은 유사한 성능을 보인다. 또한 양자화를 적용하지 않았을 때와도 근사한 성능을 나타내고 있다.

시뮬레이션 결과는 신호 대 잡음비(Signal-to-Noise Ratio: SNR)가 높아질수록 테스트에 사용할 부호어의 수를 늘려가며 측정하였다. 낮은 신호 대 잡음비 영역에서는 레벨 2, 레벨 4와 같이 양자화 비트 수의 제한을 크게 했을 때보다 양자화를 적용하지 않았을 때가 더 좋은 성능을 보인다. 그러나 높은 신호 대 잡음비 영역에서는 양자화를 적용하지 않았을 때가 항상 좋은 성능을 나타내지는 않으며, 약간 성능이 떨어지는 경향도 있다. 전체적으로는 레벨 2의 양자화를 제외하면 거의 근사한 성능 차이를 보인다.

Ⅳ. 결론

본 논문은 Neural MS 복호기의 가중치, 편향치에 적용하는 양자화 레벨을 각각 달리함으로써 나온 복호 결과를 제시한다. 레벨 2로 양자화한 경우를 제외하면 양자화를 실시하지 않았을 때와 성능 차이가 거의 나타나지 않는 것을 확인할 수 있다. 특히 높은 신호 대 잡음비 영역에서는 양자화를 하지 않았을 때보다 양자화를 한 것이 성능이 더 좋게 나오는 경우도 있었다.

본 연구의 결과는 딥러닝을 사용한 구조에서는 가중치, 편향치에 해당하는 매개변수를 양자화하더라도 성능이 크게 달라지지 않는다는 것을 제시한다. 단, 실험 결과에서 양자화 레벨이 2일 때는 각 매개변수들이 두 가지의 값만을 가지는 극단적인 양자화에 해당하기 때문에 양자화를 실시하지 않았을 때와 비교하여 뚜렷한 성능 저하를 보인다.

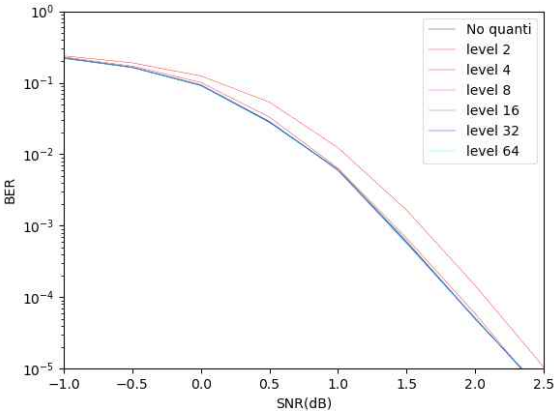

그림 2. 양자화 레벨에 따른 비트 오류율

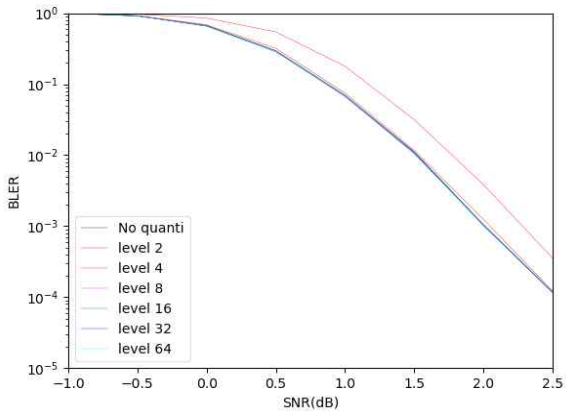

그림 3. 양자화 레벨에 따른 블록 오류율

ACKNOWLEDGMENT

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0-00846)을 받았고, 동시에 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2021-0-02068)을 받아 수행된 연구임.

참 고 문 헌

[1] Jincheng Dai, Kailin Tan, Zhongwei Si, Kai Niiu, Mingzhe Chen, H.Vincent Poor, Shunguang Cui, " Learning to Decode Protograph LDPC Codes," IEEE journal on Selected Areas in Communications, vol.39, pp. 1983-1999