딥러닝 기반 오류정정 복호기의 기술 동향 조사: 모델 프리 복호기 중심

윤홍서 [1], 곽희열 [2], *김상효 [1]
성균관대학교 전자전기컴퓨터공학과 [1],울산대학교 전기전자공학과 [2]
*iamshkim@skku.edu

Trends on Deep Learning based ECC Decoders: Focusing on Model-Free Decoders

Hong-Seo Yun[1], Hee-Youl Kwak[2], *Sang-Hyo Kim[1]
Department of Electrical and Computer Engineering, Sungkyunkwan University[1] ,University of Ulsan[2]

요 약

본 논문은 딥러닝 기반 오류 정정 부호 복호기의 동향을 분석한다. 코드의 그래프 구조를 활용하는 신경망 신뢰전파부터, 트랜스포머 모델을 적용한 ECCT 까지 주요 접근법을 검토한다. 특히 ECCT 의 연산 복잡도와 메모리 한계를 개선하기 위한 최신 경량화 및 다중 모달리티 기반 모델 프리 복호기 연구를 중심으로 기술한다.

1. 서론

차세대 통신 시스템은 초고신뢰 저지연 통신을 지원하기 위해 점점 더 복잡하고 강력한 오류 정정 부호 기술을 요구한다. 이에 최근 딥러닝 기술을 오류 정정 부호 기술에 접목하는 연구가 활발히 진행되고 있다.

초기에는 코드의 태너 그래프(Tanner graph) 구조를 신경망으로 해석한 신경망 신뢰전파 방식(Neural Belief Propagation, Neural BP) [1]이 주목받았다. 이후 기존 선형 블록 부호(Linear Block Code)의 구조에서 벗어나 모델 프리 복호를 시도하는 연구들이 시행되었고, 자연어 처리 분야에서 압도적인 성공을 거둔 트랜스포머 모델[2]이 오류정정부호 분야에 적용되어, ECCT(Error Correction Code Transformer) [3]와 같이 기존의 성능을 뛰어넘는 모델들이 등장했다.

하지만 ECCT 와 같은 모델 프리 복호기들은 우수한 성능에도 불구하고, 높은 연산 복잡도(전형적 self-attention 의 시간 복잡도는 $O(n^2d)$, 메모리 $O(n^2)$)를 갖는다[2]. 이러한 복잡도는 엄격한 지연시간(latency) 제약이 있는 환경이나 한정된 자원을 가진 엣지 디바이스에 딥러닝 복호기를 실질적으로 적용하는 데 큰 장벽으로 작용한다.

따라서 본 논문은 ECCT 와 같은 모델 프리 복호기의 실용화를 가로막는 복잡도 문제를 해결하기 위해 최근 제안되고 있는 다양한 모델 프리 경량화 아키텍처들을 중심으로 소개하고자 한다.

2. 본론

2.1 모델 기반 복호 알고리즘

모델 기반 딥러닝 복호 알고리즘은 코드의 PCM(Parity Check Matrix)나 태너 그래프와 같이 이미 알려진 구조적 정보를 신경망 설계에 직접 활용한다. 기존의 신뢰전파 알고리즘에 신경망을 적용해 개발된 신경망 신뢰전파 알고리즘에서는 기존의 계산을 신경망의 계층으로 '펼치는(unfolding)' 구조를 가진다. 즉, 태너 그래프의 변수 노드와 검사 노드 간의 메시지 반복 전달 및 업데이트 과정을 그대로 신경망의 순전파(feed-forward) 연산으로 해석하고, 각 메시지 갱신 함수에 학습 가능한 가중치(weight)를 도입하는 것이다. 신뢰전파 알고리즘과 마찬가지로, 변수 노드와 검사 노드는 LLR(Log-Likelihood Ratio) 형태의 메시지를 반복적으로 주고받으며 오류 확률을 갱신한다. 학습과정에서 가중치들이 채널 환경 및 오류 패턴에 최적화되고,결과적으로 기존 신뢰전파 방식에 비해 약 0.3~0.8 dB 의 성능 향상을 보였다([1],Fig.3, BCH(63,36),SNR = 2~4dB).

2.2 모델 프리 복호 알고리즘

2.2.1 트랜스포머

트랜스포머는 본래 자연어 처리 분야에서 제안된 모델로, 입력 시퀀스 내의 모든 요소 간 상호 관계를 병렬적으로 계산하는 Self-Attention 메커니즘을 기반으로 한다.

트랜스포머는 입력 시퀀스로부터 Q(Query), K(Key), V(Value) 행렬을 생성한다. Self-attention 메커니즘은 각 위치의 Q 벡터가 시퀀스 내 모든 위치의 K 벡터와 얼마나 연관이 있는지를 계산해 Attention Score 로 나타내고, 이 값을 V 벡터에 대한 가중치로 사용하여 문맥 정보를 종합한다. 이때 마스킹 행렬 연산(0 또는 $-\infty$ 값 적용)을 통해 인과성(causality)을 보장하거나 특정 연결을 차단할 수 있으며, 이는 정보 흐름 제어에 핵심적이다.

2.2.2 ECCT

ECCT 는 수신된 LLR 값으로 구성된 시퀀스를 입력받아, Self-Attention 메커니즘을 통해 시퀀스 내 모든 비트 위치 간의 상호 관계를 종합적으로 학습한

다. 단, 모든 코드워드를 직접 학습할 경우 코드워드 수가 2^k 에 달하여 지수적 복잡도와 과적합(overfitting) 문제가 발생한다.

ECCT 는 이 문제를 해결하기 위해 신드롬($s = Hr^T$,H 는 PCM, r 은 수신벡터) 기반 접근 방식을 채택하여, 복호기가 특정 코드워드에 불변(invariant, 선형 블록 코드의 선형성에 위배되지 않음)하도록 설계되었다. 즉, n 길이의 코드워드 시퀀스 대신, $2n-k$길이(n 길이의 수신된 LLR 벡터와 $n-k$길이의 신드롬벡터 연결)의 시퀀스 벡터([n],[n-k])를 입력으로 사용한다. 이 시퀀스는 임베딩 계층을 통해 고차원 벡터 공간으로 사상되고, 학습된 가중치 행렬을 곱해 Q,K,V 를 생성한다.

또한 PCM 으로부터 마스킹을 생성한 후에 Attention 생성에 적용해, 태너 그래프의 기본연결(비트-신드롬)과 동일 검사노드를 공유하는 비트들 간 관계(비트-비트)까지만 정보로서 반영한다. 이로써 ECCT 는 트랜스포머의 전역 표현력을 유지하면서도 코드 구조의 일관성을 보장한다.

2.3 모델 프리 알고리즘 신규 연구 동향

2.3.1 Attention 경량화 및 양자화 기법

ECCT 의 높은 연산 복잡도와 메모리 요구량을 결하기 위해 제안된 모델인 AECCT (Accelerated ECCT)[4]는 ECCT 의 핵심 복잡도 병목인 Attention 연산과 선형 레이어를 개선하기 위해 세 가지 주요 기법을 통합한다.

첫째, HPSA (Head Partitioning Self Attention) 는 기존 ECCT 의 Attention 헤드를 1-링(V-C) 연결 전담 그룹과 2-링(V-V, C-C) 연결 전담 그룹으로 명시적으로 분할하여, Attention 마스크의 희소성(sparsity)을 극대화하고 연산 복잡도를 약 40% 낮춘다([4], Table 1, BCH(63,45), GPU Inference Complexity, HPSA 적용 시 FLOPs 39.6% 감소).

둘째, AAP (Adaptive Absolute Percentile) 양자화 기법을 통해 트랜스포머 내부의 선형 레이어 가중치를 3 진(Ternary, $\{-1,0,+1\}$) 값으로 압축하여, 곱셈 연산을 제거하고 메모리 사용량을 약 70% 절감([4], Fig. 6, BCH(63,45), AAP 양자화 적용 시 Model Size 68.7% 감소)한다.

셋째, SPE (Spectral Positional Encoding) 를 도입하여 태너 그래프의 라플라시안 고유공간 정보를 위치 인코딩으로 활용함으로써, ECCT 의 이진(binary) 마스킹에서 손실되었던 세분화된 그래프 구조 정보를 모델에 제공한다.

2.3.2 다중 모달리티 아키텍처

Cross-MPT [5]는 수신된 LLR 시퀀스와 코드의 패리티 검사 행렬을 서로 다른 모달리티(modality, 서로 다른 정보 표현 형태를 의미하며, 예를 들어 이미지·텍스트·그래프 등 데이터 유형의 차이를 표현)로 간주한다. 이 모델은 입력을 패치(patch) 단위로 분할하고, Cross-Attention 메커니즘을 통해 이 두 모달리티의 상관관계를 학습한다. LLR 정보만을 사용하는 ECCT 와 달리, 코드의 구조적 정보를 함께 활용해 복잡도를 낮추고 BER(Bit Error Rate)성능을 증진한다. ([5] Fig. 4, BCH(63,36), SNR=2.5~3.5 dB, ECCT 대비 BER 약 0.35 dB 향상, 복잡도 약 25% 감소)

2.3.3 하이브리드 아키텍처

신경망 반복복호 알고리즘의 낮은 복잡도와 트랜스포머의 높은 성능(전역적 표현력)이라는 장점을 결합하려는 시도이다. Diff-MPT (Differential-Attention Message Passing Transformer)[6]는 메시지 전달 네트워크 내부에 경량 Attention 모듈을 삽입하여, 지역적(local) 그래프 메시지와 전역(global) 시퀀스 정보를 동시 활용한다. 이 접근법은 복호 성능 향상과 지연 시간 감소 사이의 균형점을 찾는 데 중요한 연구 방향으로 주목받고 있다.

3. 결론

본 논문은 딥러닝 기반 모델 프리 복호기인 ECCT 와 Cross MPT, Diff-MPT, AECCT 등의 최신 동향을 소개하였다.

본론에서 살펴본 바와 같이, 다중 모달리티 아키텍처, 하이브리드 아키텍처, 효율적 Attention 메커니즘 등 다양한 개선 기법들이 활발히 연구되고 있다.

결론적으로, 현재의 모델 프리 복호기는 직접 적용하기에는 한계가 명확하다. 향후 연구는 ECCT 의 우수한 복호 성능을 유지하면서도, 앞서 언급된 경량화, 양자화, 모달리티 융합 등의 기법들을 접목하여 연산 복잡도를 획기적으로 줄이는 방향에 중점을 두어야 할 것이다.

ACKNOWLEDGEMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2024-00398449), 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원 (No.2021-0-00794) 및 한국연구재단의 지원을 받아 수행된 기초연구사업임 (RS-2024-00343913).

4. 참고 문헌

[1] E. Nachmani *et al.*, " Learning to Decode Linear Codes Using Deep Learning," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 119-131, Feb. 2018.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., “Attention Is All You Need,” *Proc. 31st Conf. Neural Information Processing Systems (NIPS’17)*, Long Beach, CA, USA, Dec. 2017, pp. 6000-6010.

[3] M. Choukroun and L. Wolf, "Error Correction Code Transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 12046-12055.

[4] M. Levy, Y. Choukroun, "Accelerating error correction code transformers," arXiv:2410.05911, 2024.

[5] S.-J. Park *et al.*, "CROSSMPT: Cross-Attention Message Passing Transformer for Error Correcting Codes,"

[6] C. W. Lau *et al.*, "Interplay Between Belief Propagation and Transformer: Differential-Attention Message Passing Transformer," *arXiv preprint arXiv:2509.15637*, 2025.