자원 제한 환경에서의 시맨틱 통신을 위한 경량화 기법 연구

이수진, 이예훈
서울과학기술대학교

{sujinlee, y.lee}@seoultech.ac.kr

Lightweight Techniques for Semantic Communication in Resource-Constrained Environments

Su-Jin Lee, Ye Hoon Lee
Seoul National University of Science & Technology

요 약

본 논문은 자원 제한적인 네트워크 환경에서 IoT 단말과 기지국 간의 계산 용량 차이를 고려한 비대칭 구조의 시맨틱 통신 모델을 제안한다. IoT 단말의 인코더에 프루닝과 양자화 기법을 결합하여, 계산 복잡도와 메모리 사용량을 줄이면서도 성능 저하를 최소화했다. 이를 통해 6G 비지상망-IoT 네트워크에서 에너지 효율적인 시맨틱 통신의 구현 가능성을 제시한다.

Ⅰ. 서 론

시맨틱 통신은 중요한 정보만을 전송함으로써 전송 효율성을 높이고, 채널 변화에 견고한 성능을 보인다 [1][2]. 이는 6G 비지상망-IoT 네트워크 환경의 초저지연 대용량 데이터 활용 측면에서 중요한 역할을 할 것으로 기대된다. 그러나 딥러닝 기반 시맨틱 통신 모델은 높은 계산 용량을 요구하기 때문에, 상대적으로 전력과 메모리가 제한된 위성 또는 IoT 단말에서 온보드/엣지 컴퓨팅을 구현하는 데 어려움이 있다. 따라서 낮은 계산 복잡도와 모델 경량화가 필수적이다.

딥러닝 모델 경량화 기법으로 프루닝과 양자화 기법이 널리 사용된다 [3]. 가중치 프루닝은 사전 학습된 모델에서 불필요한 가중치를 제거하여 파라미터 개수를 줄이고, 양자화는 가중치를 더 적은 비트 수로 표현하여 메모리 사용을 줄인다.

본 논문에서는 자원 제한 환경에서 시맨틱 통신을 구현하기 위해 프루닝과 양자화 기법을 활용한 자원 효율적인 모델 경량화 방식을 제안한다.

Ⅱ. 시스템 모델

본 시스템은 IoT 네트워크에서 IoT 단말과 기지국 간 상향링크를 고려한다. 오토인코더 기반 시맨틱 통신 모델에서 IoT 단말과 기지국의 계산 용량 차이를 반영하여, IoT 단말 내의 인코더 모델만 압축하는 비대칭 구조를 따른다. 이 구조는 IoT 단말의 제한된 자원을 효율적으로 사용하면서, 기지국 측에서는 원본 모델의 성능을 유지할 수 있다. 인코더의 모델 압축 비율은 가중치 프루닝 비율과 양자화 비트 수에 따라 결정된다. 가중치 프루닝은 모델 가중치의 크기를 기반으로 중요도를 판단하여, 정해진 임계값 이하의 가중치를 제거한다. 이후 프루닝된 모델의 가중치는 선형 양자화를 통해 FP32 에서 m비트 정수로 압축된다.

Ⅲ. 실험 결과

본 연구는 Rician 페이딩 채널을 가정하며, 해당 채널은 $\mathcal{CN}(\mu, \sigma^2)$ 분포를 따른다. 여기서 μ는 $\sqrt{k/(k+1)}$, σ는 $\sqrt{1/(k+1)}$이며, Rician 계수 k 는 1 로 설정했다. 시맨틱 통신 모델은 트랜스포머 기반의 L-DeepSC [4] 프레임워크를 사용하였다. 텍스트 기반 시맨틱 통신의 성능 평가에는 BLEU 점수를 활용하였다.

Ⅲ-1. 비대칭 오토인코더 구조

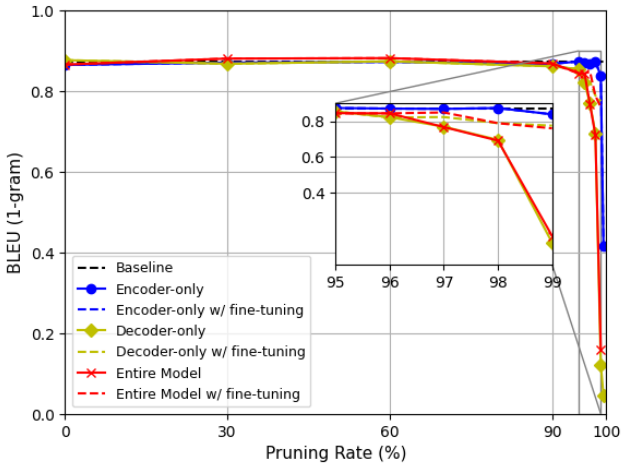

그림 1. 다양한 가중치 프루닝 모델의 프루닝 비율에 따른 BLEU 점수 비교

그림 1 은 비대칭 구조와 대칭 구조의 프루닝 비율에 따른 BLEU 점수 비교를 나타낸다. Baseline 은 프루닝하지 않은 원본 모델로, BLEU 점수가 0.8726 으로 일정하다. 실험 결과, 디코더만 프루닝한 모델과 모델 전체를 프루닝한 대칭 구조 모델의 성능은 유사한 반면, 인코더만 프루닝한 비대칭 구조 모델은 프루닝 비율이 90% 이상일 때 더 나은 성능을 보였다. 이는 IoT-기지국 상향링크에서 기지국의 컴퓨팅 자원이 충분한 경우, 비대칭 구조를 통해 효과적으로 IoT 단말의 모델을 경량화 할 수 있음을 나타낸다. 또한 인코더만 프루닝한 모델에서 성능 저하 없이 98%까지 프루닝이 가능함을 확인했다. 디코더 또는 전체 프루닝 모델은 파인튜닝을 통해 성능을 개선할 수 있다.

III-2. 프루닝-양자화 결합 모델

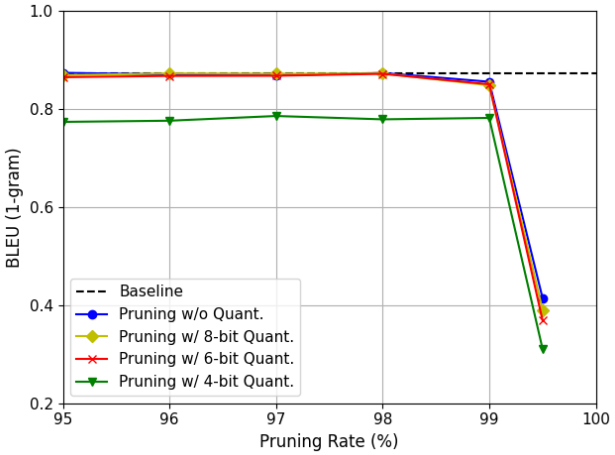

그림 2. 인코더 프루닝 모델의 양자화 결과 비교

표 Ⅰ. 모델 경량화 결과

	#Parameters	Size	BLEU score
$\gamma = 0$, $m = 32$	823,296	3.29MB	0.8726
$\gamma = 0.98$, $m = 8$	16,465	0.16MB	0.8723

그림 2 는 인코더 프루닝 모델에 양자화 적용 여부에 따른 BLEU 점수의 변화를 나타내며, 6 비트 양자화까지는 성능 저하가 거의 없음을 보인다. 실제로는 하드웨어 호환성 문제와 구현 복잡성으로 인해 8 비트 양자화가 빈번하게 사용되며, 이는 모델의 성능 저하를 거의 유발하지 않으면서 모델 압축과 정확도 사이에서 최적의 균형을 제공한다.

표 Ⅰ은 프루닝 비율 γ와 비트 수 m에 따른 인코더 모델 크기의 감소를 보여준다. 가중치 프루닝과 양자화의 결합을 통해 모델 크기가 3.29MB 에서 0.16MB 로 감소하며, BLEU 점수는 유사한 수준을 유지한다. 이를 통해 경량화된 시맨틱 통신 모델이 제한된 하드웨어 자원에서도 효율적으로 동작할 수 있음을 보여준다.

Ⅳ. 결론

본 논문에서는 자원 제한 환경에서 시맨틱 통신을 위한 비대칭 구조의 모델 경량화 기법을 제안하였으며, 실험을 통해 인코더 프루닝과 양자화를 결합한 방식이 성능 저하 없이 98%까지 모델 압축이 가능함을 확인하였다. 이 연구는 자원 제한적인 비지상망-IoT 네트워크에서 효율적인 시맨틱 통신 구현을 위한 가능성을 제시하며, 특히 프루닝과 양자화를 결합하여 모델의 경량화와 성능 사이의 균형을 효과적으로 달성할 수 있음을 보여준다. 향후 연구에서는 다수의 IoT 단말과 위성 통신 환경에서 각 단말과 전체 네트워크의 성능과 에너지 효율 간의 균형을 맞추기 위한 최적화 기반 프루닝 전략을 연구할 예정이다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2021-0-00368, 6G 서비스를 위한 인공지능/머신러닝 기반 자율형 MAC 개발)을 받아 수행된 연구임.

참 고 문 헌

[1] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567 579, 2019.

[2] H. Xie, Z. Qin, G. Y. Li, and B. -H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663-2675, 2021.

[3] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370-403, 2021.

[4] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142-153, Jan. 2021.