트랜스포머 기반 시맨틱 통신을 위한 모델 경량화 및 시스템 최적화 기법 조사

반동현, 서효운*
성균관대학교, 성균관대학교

dhban@skku.edu, * hyowoonseo@skku.edu

A Survey on Lightweight and System Optimization Techniques for Transformer-based Semantic Communication

Donghyeon Ban, Hyowoon Seo*
Sungkyunkwan Univ., *Sungkyunkwan Univ.

요 약

본 논문은 트랜스포머 기반 시맨틱 통신의 최적화를 위한 핵심 기술들을 모델 가중치 압축, 시맨틱 표현 수준 압축, 지식 증류, 시스템 관점 최적화의 네 가지 범주로 분류하고, 각 기법의 적용 사례와 성능 특성을 비교 분석하였다. 또한 향후 시맨틱 통신 시스템 설계에 미치는 영향을 파악하기 위하여 다양한 기법들의 통합적 연구의 필요성을 논의하였다.

Ⅰ. 서 론

시맨틱 통신(Semantic Communication)은 정보의 '의미적 전달'을 목표로 하는 통신 시스템이다. 이 시스템에서 송신기는 데이터의 의미론적 표현만을 추출하여 전송하고, 수신기는 해당 표현을 바탕으로 원래 정보를 재구성한다. 의미론적 표현만을 전송하면 비트 단위 무결성을 전제로 한 기존 통신 시스템의 전송 비효율을 제거할 수 있어, 전송 효율과 통신 신뢰도를 동시에 향상시킬 수 있는 방식으로 주목받고 있다 [1]. 시맨틱 통신의 '의미 추출-복원' 단계에서 딥러닝 기반 트랜스포머(Transformer)가 핵심 엔진으로 널리 채택되는데, 이는 self-attention 메커니즘 덕분에 텍스트·이미지·음성 등 다양한 포맷에 적용 가능하고 데이터 내재 정보의 장거리 의존성을 정밀하게 학습할 수 있기 때문이다 [2].

그러나 대규모 트랜스포머 모델은 높은 연산 복잡도와 큰 메모리 용량을 요구한다. 이러한 요구 사항은 시맨틱 통신을 다양한 환경에 적용하는 데 큰 제약으로 작용한다. 실시간 처리를 위해서는 모델 크기를 경량화하고, 통신 상태에 따라 모델을 적응적으로 활용할 수 있도록 시스템을 최적화할 필요가 있다.

본 논문은 트랜스포머 기반 시맨틱 통신의 최적화 핵심 기술을 모델 가중치 압축(Weight-Level Compression), 표현 수준 압축(Semantic Representation Compression), 지식 증류(Knowledge Distillation)를 활용한 모델 축소, 시스템 관점 최적화의 네 가지 범주로 조사하였다. 실제 연구들은 이들 기술을 복합적으로 결합하여 자원을 절감하면서도 의미 정보를 보존하고 있으며, 본 논문은 각 기법의 원리와 무선 시뮬레이션 결과를 토대로 이러한 성과를 비교·분석하였다.

Ⅱ. 본론

모델 가중치 압축(Weight-Level Compression)은 부동소수점 파라미터를 저비트 정수로 치환해 모델 파일 크기와 메모리 사용량을 동시에 줄인다. IAQ [3]는 Vision Transformer 의 어텐션 점수를 이용해 '가중치 양자화 오차 최소화' 문제를 볼록 최적화 문제로 정식화하고, 오차 감소 폭이 가장 큰 패치에 비트를 순차 할당하는 incremental allocation 과 KKT 기반 워터필링 해석식을 제안하였다. 실험 결과, 양자화 압축률 ρ = 0.125 에서도 MIRO 데이터셋 정확도 93 % 이상을 유지하며 Fixed-Q 및 Top-k 대비 5– 8 %p 높은 성능을 달성하였다.

시맨틱 표현 수준 압축(Semantic Representation-Level Compression)은 인코더가 생성한 시맨틱 표현을 양자화·부호화해 전송 효율과 의미 보존을 동시에 확보한다. GOS-VAE 는 송신기에 경량 VQ-VAE 인코더를, 수신기에 디코더와 OneFormer 를 분할 배치하여 r = 4 에서 mIoU 57.3 %를 기록하고, 기본 VQ-VAE 대비 4.3 %p 성능 향상과 파라미터 수 92 배 절감을 동시에 달성하였다. 이러한 코드북 최적화 및 의미 중심 비트 할당은 동일 대역폭에서 PSNR 을 3– 4 dB 높이고 시각적 품질도 향상시킴을 보여준다 [5]. uJSCC 는 시맨틱 특징을 정수 인덱스로 매핑하는 벡터 양자화를 통해 아날로그적 JSCC 를 기존 디지털 PHY 와 자연스럽게 호환시키는 범용 (Universal) 접근을 제시하였다. [4]

지식 증류(Knowledge Distillation) 기반 모델 축소는 대규모 교사 모델의 지식을 소형 학생 모델로 이전해 경량화하면서도 성능을 유지한다. KD-MU-SemCom [7]은 Swin-Transformer 디코더(teacher)와 경량 디코더(student)를 FRENCA-KD 절차로 학습하여 AWGN SNR = 3 dB 환경에서 PSNR 을 0.68 dB, MS-SSIM 을 0.08 만큼 향상시키고 연산량을 17 % 절감하였다. KD-SC-MultiUser [8]는 DeepSC 계층을

교사로 증류한 학생 모델이 SNR < 9 dB 구간에서 BLEU 점수를 기준선보다 10 %p 이상 높이고 파라미터를 약 44 % 줄였다. 또한 FSSC [9]는 연합 학습, 모델 슬라이싱, KD 를 결합해 단말 모델 크기를 0.05 MB 까지 축소하면서도 PSNR 을 추가로 2 dB 개선하였다.

시스템 관점 최적화(System-Level Optimization)는 모델 구조와 연산 위치를 조정해 엔드 투 엔드 자원 효율을 극대화한다. Swin-Transformer Semantic Communication(STSC)은 윈도우 어텐션과 두 차례 Patch Merging 을 통해 연산량을 $O(N^2)$에서 $O(N)$으로, 토큰 수를 1/16 로 줄여 압축률 0.33 에서 CNN-JSCC 대비 평균 PSNR 을 5 dB 향상시켰다. 이를 연합 학습으로 확장한 Federated STSC(FSSC)는 중앙 데이터 이동 없이 FedAvg 로 60 round 이내에 수렴하여 중앙집중식 대비 MSE 를 10 % 낮추고(PSNR 2– 3 dB 증가) 서버 부하도 분산시켰다 [9]. 이러한 모델·채널 양면 경량화와 학습·추론 분산 기법을 통합 적용하면, 실시간 시맨틱 통신에서도 자원 절감과 품질 보존을 동시에 확보할 수 있음을 확인하였다.

표 1 은 앞서 소개한 논문들의 세부 기법과 주요 성과를 요약한 것이다.

표 1 경량화 기법 별 주요 성과

경량화 기법	세부 기법	주요 성과
모델 가중치 압축	중요도 가중 비트 할당 양자화	동일 정확도 기준 전송량 20 – 40 % 추가 절감 동일 ρ 기준 Top-1 정확도 향상 [3]
	어텐션 유도 다중 해상도, 다단계 양자화	단일 해상도 대비 고압축 환경에서 15–60 %p 정확도 향상 [6]
시맨틱 표현 수준 압축	VQ-VAE 기반 벡터 양자화, 모듈 경량화	압축률 r=4 환경 모델 크기 99% 감소 mIo+ 2.3p, 전송대역폭 −4.4 KB. [5]
지식 증류(KD)	KD + 전이학습	기존 반복학습 대비 PSNR +1.25 dB / MS-SSIM +0.012. [7]
	KD + 후처리 동적 양자화	Teacher 모델 대비 Student 2: 모델 크기 44% 감소, 저 SNR 상황 BLEU 10 %p 향상 [8]
시스템 관점 최적화	System-Level Optimization	JSCC 대비 STSC: 평균 5 dB PSNR 향상[9]

Ⅲ. 결론

본 논문은 트랜스포머 기반 시맨틱 통신 모델의 성능과 자원 효율성을 동시에 향상시키기 위한 최신 경량화 기술을 모델 가중치 압축, 시맨틱 표현 수준 압축, 지식 이전 기반 모델 축소, 시스템 관점 최적화의 네 범주로 분류하고, 연구 사례를 비교·분석하였다. 분석 결과, 토큰 중요도 기반 벡터 양자화, 교사–학생 지식 증류, 분할 추론 및 적응형 전송 등 여러 기법을 복합 적용한 전략이 실시간 환경에서도 뛰어난 자원 절감 효과와 성능 우위를 달성함을 확인하였다. 앞으로는 각 단일 기법과 이들의 융합 방식이 통신 시스템 성능에 미치는 영향을 정량적으로 평가하는 연구가 필요하며, 이를 통해 시맨틱 통신 시스템 설계에 최적화 기법을 체계적으로 적용할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 한국연구재단을 통해 과학기술정보통신부의 「한-핀란드 공동연구사업」의 지원을 받아 수행되었음(RS-2024-00464570)

참 고 문 헌

[1] H. Seo, J. Park, M. Bennis, and M. Debbah, “Semantics-Native Communication via Contextual Reasoning,” IEEE Transactions on Cognitive Communications and Networking, vol. 9, no. 3, pp. 604– 617, Jun. 2023.

[2] Y. Wang et al., "Transformer-Empowered 6G Intelligent Networks: From Massive MIMO Processing to Semantic Communication," IEEE Wireless Commun., vol. 30, no. 6, pp. 127– 135, Dec. 2023.

[3] J. Park et al., "Vision Transformer-based Semantic Communications with Importance-Aware Quantization," arXiv:2401.01234, 2024.

[4] Y. Huh, H. Seo, and W. Choi, “Universal Joint Source-Channel Coding for Modulation-Agnostic Semantic Communication,” IEEE Journal on Selected Areas in Communications, early access, doi: 10.1109/JSAC.2025.3559138, 2025.

[5] Y.-C. Chao et al., "Task-Driven Semantic Quantization and Imitation Learning for Goal-Oriented Communications," arXiv:2503.04567, 2025.

[6] M. Mortaheb, "Efficient Semantic Communication through Transformer-Aided Compression," arXiv:2405.06789, 2024.

[7] L. Nguyen et al., "Optimizing Multi-User Semantic Communication via Transfer Learning and Knowledge Distillation," IEEE Commun. Letters, vol. 28, no. 2, pp. 345– 349, 2024.

[8] C. Liu, Y., "Knowledge Distillation-Based Semantic Communications for Multiple Users," IEEE Trans. Wireless Commun., vol. 22, no. 11, pp. 7025– 7039, 2023.

[9] Y. Yan et al., "FSSC: Federated Learning of Transformer Neural Networks for Semantic Image Communication," arXiv:2504.09876, 2025.