시맨틱 통신과 생성형 인공지능 융합에 관한 최신 연구 동향

이승찬, 이동현, 송치현, 김재민, 조성래
중앙대학교 컴퓨터공학과

{sclee, dhlee, chsong, jmkim}@uclab.re.kr, srcho@cau.ac.kr

Recent Advances in Semantic Communication and Generative AI Fusion Systems

Seungchan Lee, Donghyun Lee, Chihyun Song, Jaemin Kim and Sungrae Cho
Dept. of Computer Science and Engineering
Chung-Ang University

요 약

시맨틱 통신은 송수신 간에 데이터의 의미만을 효율적으로 전달함으로써 차세대 통신 환경의 대용량 트래픽 문제를 해결하고자 등장한 패러다임이다. 기존 통신이 비트 정확도에 초점을 맞춘 반면 시맨틱 통신은 의미 전달을 중시하나, 현재의 딥러닝 기반 구현에는 한정된 일반화 성능과 낮은 문맥 이해 능력 등의 한계가 있다. 최근 부상한 생성형 인공지능 기술(대규모 언어 모델, 멀티모달 생성 모델 등)은 사람 수준의 맥락 추론과 콘텐츠 생성 능력을 갖추고 있어 이러한 한계를 보완할 수 있는 대안으로 주목받고 있다. 본 논문은 시맨틱 통신과 생성형 AI 의 융합에 관한 최신 연구 동향을 서베이하고, 기술 융합으로 인한 통신 방식의 변화, 주요 응용 사례, 그리고 향후 해결해야 할 과제를 논의한다.

Ⅰ. 서 론

초연결·초지능화를 지향하는 6G 시대를 앞두고, 시맨틱 통신은 단순한 비트 정확도보다 데이터의 의미 전달을 우선시하는 새로운 패러다임으로 주목받고 있다[1][2]. 송신단은 원본 데이터에서 불필요한 정보를 제거하고 핵심 의미만 인코딩하며, 수신단은 공유된 지식을 활용해 의미를 복원한다. 이를 통해 일정 수준의 오류가 발생하더라도 의미 이해에는 문제가 없으며, 전송 효율을 극대화할 수 있다. 그러나 현존하는 딥러닝 기반 시맨틱 통신은 특정 작업에 최적화되어 있어 일반화 성능과 문맥 추론 능력이 부족하다는 한계가 있다[3]. 또한 방대한 사전 지식 데이터 구축이 필수적이라는 점에서 비용과 시간이 과다하게 소요된다. 이러한 한계 극복을 위해 최근 생성형 인공지능(Generative AI)이 주목받고 있으며, 대규모 언어 모델(LLM)과 멀티모달 생성 모델은 인간 수준의 문맥 이해와 콘텐츠 생성 능력을 통해 시맨틱 통신 발전에 새로운 전기를 마련하고 있다.

Ⅱ. 본론

생성형 AI 융합에 따른 시맨틱 통신의 진화: 생성형 AI 를 시맨틱 통신에 활용함으로써 얻을 수 있는 대표적 이점은 다음과 같다. 첫째, 대규모 생성 AI 모델을 통해 방대한 학습 데이터와 배경 지식을 자동 생성함으로써 시맨틱 인코더/디코더의 훈련을 효율화할 수 있다[4]. 실제로 생성 AI 가 만든 다양한 AI 생성 콘텐츠(AIGC)를 시맨틱 통신 모델 학습 및 지식 베이스 구축에 활용하면, 별도의 데이터 수집 없이도 모델의 성능을 향상시킬 수 있다. 둘째, 생성 모델의 고도화된 문맥 추론 능력은 시맨틱 복원 정확도를 높여준다. 예를 들어 프롬프트 엔지니어링 기법을 통해 생성 모델이 맥락적으로 정확한 콘텐츠를 만들어내도록 함으로써, 전송 중 발생하는 잡음이나 오류에도 의미 왜곡이 최소화되도록 할 수 있다. 셋째, 프롬프트 전달을 통한 정보 재생성이라는 새로운 통신 방식이 가능해진다. 잘 학습된 생성형 모델을 송수신단에 갖춘 경우, 송신단은 원본 데이터를 요약한 간단한 프롬프트만을 전송하고 수신단은 이를 기반으로 원본과 동등한 품질의 데이터를 현지에서 생성할 수 있다. 이를 통해 모든 원본 정보를 보내지 않고도 수신 측에서 필요한 내용을 복원하거나 생성하게 되어 전송 데이터량을 획기적으로 줄일 수 있다. [5]에서는 이러한 아이디어를 구체화하기 위해 정보 보틀넥 이론과 확률적 생성 모델을 접목한 시맨틱 통신 기법을 제안하였는데, 채널 상태에 따라 전송할 정보량을 조절하고 수신단에서 원본 데이터를 재생성함으로써 주어진 전송 지연 및 전력 예산 내에서 통신 효율을 극대화할 수 있다고 보고하였다. 이처럼 생성형 AI 의 도입은 시맨틱 통신의 전송 효율 향상, 의미 전달 정확도 제고, 망 자원 절감 등으로 기존 한계를 극복할 수 있는 새로운 가능성을 열어주고 있다.

융합 시스템의 주요 응용 사례: 시맨틱 통신과 생성형 AI 의 결합은 다양한 분야의 응용에서 혁신적인 성능 향상을 가져올 것으로 기대된다. [6]에서는 생성형 AI 기반 Generative SemCom 시스템을 설계하고, 대용량 영상 데이터 전송 시 실제 효과를 평가하였다. 그 결과, 제안된 LLM 기반 생성형 통신 기법을 활용하면 전통적 통신 대비 전송해야 할 비트 수를 99.98%까지 감소시키면서도 정보 검색 정확도는 53% 향상시킬 수 있음이 보고되었다. 이는 수신기에서 의미를 재생성(regeneration)함으로써 통신 부하를 획기적으로

줄인 대표적인 사례이다. 이러한 생성형 시맨틱 통신은 산업 IoT 의 센서 데이터 전달, 자율주행 차량(V2X) 간 주변 상황 공유, 메타버스 환경에서의 몰입형 콘텐츠 전송, 드론과 같은 무인이동체 영상 스트리밍 등 다양한 영역에 적용될 수 있을 것으로 전망된다. 예를 들어, 자율주행차 네트워크에서는 차량에 장착된 로컬 LLM 기반 에이전트가 각종 센서 데이터를 실시간 해석하여 도로 상황을 요약 전송하고, 수신 차량은 이를 토대로 필요한 경고 또는 제어 신호를 자체 생성함으로써 초저지연 협력 운행을 구현할 수 있다[7]. 메타버스 기기 간 통신의 경우에도 사용자 동작을 의미 묘사로 추출하여 주고받고, 수신 단말이 해당 묘사에 따라 가상 아바타나 환경을 즉각 생성함으로써 대역폭 부담 없이도 실시간 상호작용을 가능케 한다. 이처럼 여러 응용 시나리오에서 시맨틱 통신+생성형 AI 융합은 적은 전송으로 풍부한 결과를 얻는 새로운 통신 서비스를 제공할 수 있다.

향후 연구 과제: 생성형 AI 와의 융합을 현실의 통신망에 적용하기 위해서는 해결되어야 할 기술적 과제들도 남아 있다. 첫째로, GPT-3 와 같은 초거대 모델은 수십억 이상의 파라미터로 이루어져 있어 동작에 막대한 연산 자원과 저장 공간을 필요로 하므로 이러한 모델을 엣지 단말에서 효율적으로 구동하기 위한 경량화 및 분산 처리 기술이 요구된다. 둘째로, 생성 AI 출력의 신뢰성에 관한 문제이다. 생성형 모델이 산출하는 콘텐츠는 자동 생성의 특성상 일정 수준의 불확실성을 내포하며, 추가적인 지연을 발생시킬 우려도 있다. 통신 맥락에서 이러한 불확실성과 지연은 신뢰성과 지연 민감도 측면의 새로운 도전과제가 되므로, 생성 모델의 출력 품질 보장 및 실시간 처리에 대한 연구가 필요하다. 셋째로, 프라이버시 및 보안 이슈가 있다. 시맨틱 통신은 송수신단이 배경 지식을 공유하는 것이 전제되는데, 개인 데이터나 선호도 등의 민감 정보를 활용하면서도 프라이버시를 보호할 수 있는 방안이 모색되어야 한다. 아울러 사용자들이 자신의 데이터를 학습에 제공하거나 배경지식으로 공유하려 하지 않을 경우 시스템 성능이 제한될 수 있으므로, 인센티브 메커니즘 등의 정책적 지원 방안도 고려되어야 한다. 이외에도 생성 모델과 통신 시스템 간의 표준 인터페이스 정립, 오류 발생 시 의미 왜곡을 방지하는 향상된 부호화 기법, 그리고 평가 기준 마련 등 다양한 후속 연구가 제기되고 있다.

Ⅲ. 결론

본 논문에서는 차세대 통신 패러다임으로 주목받는 시맨틱 통신과 생성형 인공지능 기술의 융합에 대한 연구 동향을 살펴보았다. 시맨틱 통신은 의미 기반의 효율적 정보 전달을 가능하게 하나, 현존 기법들의 한계로 인해 추가적인 지능화가 요구된다. 이러한 배경에서 등장한 생성형 AI 는 방대한 데이터 학습으로 얻은 생성 능력과 추론 능력을 바탕으로 시맨틱 통신의 성능 향상과 새로운 서비스 구현에 기여하고 있다. 여러 응용 분야에서 의미 있는 성과가 보고되고 있으며, 남아 있는 도전 과제들에 대해서도 활발한 연구가 진행 중이다. 향후 생성형 AI 와 통신기술의 긴밀한 융합을 통해 고효율·지능형 통신 시대가 앞당겨질 것으로 기대된다.

ACKNOWLEDGMENT

This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation) - ITRC (Information Technology Research Center) (IITP-2025-RS-2022-00156353, 50% / IITP-2025-RS-2024-00436887, 50%) grants funded by the Korea government (Ministry of Science and ICT).

참 고 문 헌

[1] S. Park, C. Park and J. Kim, "Learning-Based Cooperative Mobility Control for Autonomous Drone-Delivery," in *IEEE Transactions on Vehicular Technology*, vol. 73, no. 4, pp. 4870-4885, April 2024.

[2] W. J. Yun, D. Kwon, M. Choi, J. Kim, G. Caire and A. F. Molisch, "Quality-Aware Deep Reinforcement Learning for Streaming in Infrastructure-Assisted Connected Vehicles," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2002-2017, Feb. 2022.

[3] S. Park, J. P. Kim, C. Park, S. Jung and J. Kim, "Quantum Multi-Agent Reinforcement Learning for Autonomous Mobility Cooperation," *IEEE Communications Magazine*, vol. 62, no. 6, pp. 106-112, June 2024.

[4] L. Xia et al., "Generative AI for Semantic Communication: Architecture, Challenges, and Outlook," *arXiv* preprint arXiv:2308.15483, 2024.

[5] S. Barbarossa *et al.*, "Semantic Communications Based on Adaptive Generative Models and Information Bottleneck," *IEEE Communications Magazine*, vol. 61, no. 11, pp. 36- 41, 2023.

[6] J. Ren *et al.*, "Generative Semantic Communication: Architectures, Technologies, and Applications," *arXiv* preprint arXiv:2412.08642, 2024.

[7] Z. Wang *et al.*, "Large-Language-Model-Enabled Text Semantic Communication Systems," *Applied Sciences*, vol. 15, no. 13, Article 7227, 2025.