시맨틱 통신과 AI 기반 자원 최적화 동향

이승찬, 이충현, 김가현, 조성래
중앙대학교 컴퓨터공학과

{sclee, chlee, ghkim}@uclab.re.kr, srcho@cau.ac.kr

A Survey on State-of-the-Art Semantic Communication and AI-Based Resource Optimization

Seungchan Lee, Chunghyun Lee, Gahyun Kim and Sungrae Cho
Department of Computer Science and Engineering, Chung-Ang University

요 약

본 논문은 메시지의 핵심 의미를 인코딩·전송하여 대역폭·전력 사용량을 크게 절감할 수 있는 시맨틱 통신과, 이를 지능적으로 운용하기 위한 AI 기반 자원 최적화 기법의 동향을 조사한다. 딥러닝 모델을 활용해 텍스트·음성·영상 데이터를 의미 벡터로 변환하고, 강화학습 등을 이용해 네트워크 자원을 최적 배분함으로써 전송 효율과 QoS 를 동시에 높일 수 있음을 확인한다. 이러한 접근은 6G 시대에 요구되는 초저지연·초연결 환경을 구현할 핵심 패러다임으로 주목받고 있다.

Ⅰ. 서 론

최근 차세대 이동통신(6G) 기술에 대한 연구가 활발히 진행됨에 따라, 초저지연(Ultra-Low Latency)·초광대역(Ultra-Broadband)·대규모 사물인터넷(Massive IoT) 등 다양한 서비스 요구사항을 동시에 만족시키는 네트워크 자원 관리 전략이 중요한 연구 과제로 부상하고 있다[1][2]. 이와 함께 부각된 시맨틱 통신(Semantic Communication) 패러다임은 전송 데이터의 '의미(semantic)'를 추출하여 압축·전송함으로써, 기존 비트 단위 전송 방식에 비해 대역폭 및 에너지 효율을 획기적으로 개선할 수 있다는 가능성을 보여준다[3]. 특히, 딥러닝(Deep Learning) 기반 모델을 활용하면 텍스트·음성·영상 등 다양한 비정형 데이터를 시맨틱 레벨에서 처리하여 높은 전송 효율과 데이터 정확도를 동시에 확보할 수 있다.

한편, 멀티 엑세스 엣지 컴퓨팅(Multi-Access Edge Computing, MEC) 환경에서 AI 기반 자원 최적화 기법을 결합하면, 사용자 요구 변화나 네트워크 상태 변동에도 빠르게 대응이 가능해진다[4][5]. 예컨대 강화학습(Reinforcement Learning)은 실시간 정책 업데이트를 통해 전송 대역폭 할당, 전력 관리 등 다양한 자원 관리 문제에 효과적으로 적용될 수 있으며, 이를 시맨틱 통신과 연계할 경우 전체 시스템 효율성이 크게 향상된다는 연구 결과들이 보고되고 있다[6]. 본 논문은 이러한 시맨틱 통신과 AI 기반 자원 최적화의 최신 연구 동향을 조사·정리함으로써, 해당 분야의 발전 경향을 파악하고자 한다.

Ⅱ. 본론

시맨틱 통신(Semantic Communication)은 6G 시대에 초저지연(Ultra-Low Latency), 초광대역(Ultra-Broadband), 초연결(Massive Connectivity) 요구사항을 충족하기 위해 부상한 새로운 전송 패러다임으로, 텍스트·음성·영상 등의 데이터를 비트 단위로 전송하기보다는 핵심 의미(semantic)만을 인코딩하여 효율적으로 전달한다는 특징을 가진다. 예컨대, 음성 데이터를 처리할 때 딥러닝 기반 모델(Transformer, RNN 등)을 사용하여 발화자의 의도나 문맥을 압축된 벡터 형태로 추출하고, 이를 전송한 뒤 수신 측에서 복원함으로써 불필요한 비트 전송을 크게 줄일 수 있다[3]. 이 과정에서 시맨틱 인코딩과 디코딩 사이의 의미적 거리를 측정하기 위해, 시맨틱 왜곡(Semantic Distortion)D_{sem} 이 사용되며 수식 1 과 같다.

$$D_{sem} = \mathbb{E}[d(s, \hat{s})] \quad (1)$$

여기서 s와 $\hat{s}$는 각각 송신·수신 측의 의미 벡터이고, $d(\cdot)$는 코사인 거리 혹은 크로스 엔트로피 등 의미적 유사도를 측정하기 위한 함수이다. 이를 최적화함으로써, 전송 효율과 의미 보존도를 균형 있게 유지할 수 있으며, 최근에는 멀티미디어 스트리밍, AR/VR, 음성 비서 서비스 등에서 20~30% 이상의 대역폭 절감을 보고하고 있다[5].

한편, 시맨틱 통신이 제공하는 잠재적 이점을 현실화하기 위해서는, AI 기반 자원 최적화 기법이 필수적으로 요구된다 [6]. 제한된 네트워크 자원(대역폭, 전송 파워, 엣지 서버의 연산 능력 등)을 어떻게 할당할지 결정해야 하는 복잡한 문제에 대해, 머신러닝(ML)과 강화학습(RL)을 적용함으로써 실시간 예측과 동적 의사결정이 가능해진다. 예컨대,

에이전트(Agent)가 시맨틱 왜곡과 전송 지연(혹은 비용)을 동시에 최소화할 수 있도록 보상 함수를 설계하면, 네트워크 상태 변화에 빠르게 대응하고, 사용자별 QoS 를 맞춤형으로 보장할 수 있다는 연구가 다수 제시된다[3][4]. 특히, 멀티 엑세스 엣지 컴퓨팅(MEC) 환경에서 엣지 노드가 데이터 일부를 사전에 시맨틱 분석하고, 분산 학습(Federated Learning)을 병행하여 모델 파라미터만 공유함으로써 프라이버시 이슈와 통신 부하를 동시에 완화하는 접근이 주목받고 있다[6]. 이처럼 시맨틱 통신과 AI 자원 최적화의 결합은 이종(heterogeneous) 네트워크 구성에서도 스펙트럼 효율과 에너지 효율을 동시에 높일 수 있어, 자율주행·스마트 팩토리·원격의료 등 초실시간·초신뢰성이 요구되는 분야에 핵심 솔루션으로 부상할 전망이다.

아울러, 시맨틱 통신의 적용 범위가 확장될수록, 표준화와 상호운용성(Interoperability) 이슈가 뒤따른다는 점에도 주목할 필요가 있다. 각 서비스나 디바이스마다 요구하는 의미 수준(semantic level)이 다르고, 데이터 종류·형식에 따라 인코딩 스키마가 달라지므로, 시맨틱 계층(semantic layer)의 정의와 프로토콜 설계가 중요 과제로 남아 있다. 또한, 대규모 상용 환경에서 시맨틱 통신과 AI 최적화 모델을 운영하기 위해서는, 보안·프라이버시·지연 보장 등이 우선적으로 해결되어야 하며, 강화학습 에이전트가 대규모 트래픽과 빠르게 변동하는 사용자 요구에 대응할 수 있는 고성능 하드웨어 및 효율적인 분산 학습 프레임워크도 요구된다. 그럼에도 불구하고, 시맨틱 통신은 기존 비트 기반 통신 패러다임의 한계를 근본적으로 극복할 수 있는 잠재력을 지니고 있으며, AI 기반 자원 최적화와 결합됨으로써 6G 시대의 초연결·초지능 인프라 구축을 가속화할 것으로 기대된다.

Ⅲ. 결론

본 논문은 시맨틱 통신(Semantic Communication)이 차세대 이동통신(6G) 환경에서 기존 전송 방식의 한계를 극복할 유망한 패러다임임을 살펴보고, 이를 AI 기반 자원 최적화 기법과 결합하는 최근 동향을 간략히 조사하였다. 시맨틱 통신을 통해 텍스트·음성·영상 등 다양한 형태의 데이터를 '의미(semantic)' 수준에서 압축·전송함으로써, 대역폭·전력 사용량을 절감하면서도 높은 서비스 품질(QoS)을 유지할 수 있음이 여러 연구를 통해 제시되고 있다. 특히, 강화학습(Reinforcement Learning)이나 분산학습(Federated Learning) 등 AI 기술이 적용되면서, 동적인 네트워크 상태 변화와 사용자 요구에도 유연하게 대처할 수 있는 가능성이 한층 높아졌다.

다만, 대규모 상용 환경에서의 실증 연구와 표준화 작업은 아직 초기 단계에 머무르고 있으며, 보안·프라이버시 이슈나 상호운용성(Interoperability) 확보 등 해결해야 할 과제도 적지 않다. 그럼에도 불구하고 시맨틱 통신과 AI 최적화의 융합은 초저지연·초광대역·초연결이 필수적인 6G 시대에 핵심적 역할을 할 것으로 기대되며, 향후에도 이 분야의 연구가 더욱 활발히 진행될 것으로 전망한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구 센터육성지원사업(IITP-2025-RS-2022-00156353) 및 과학기술정보통신부의 재원으로 한국연구재단(RS-2023-00209125)의 지원을 받아 수행된 연구임

참 고 문 헌

[1] Q. Lan *et al.*, "What is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence," in *Journal of Communications and Information Networks*, vol. 6, no. 4, pp. 336-371, Dec. 2021

[2] X. Luo, H. -H. Chen and Q. Guo, "Semantic Communications: Overview, Open Issues, and Future Research Directions," in *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210-219, February 2022

[3] H. Xie, Z. Qin, G. Y. Li and B. -H. Juang, "Deep Learning Enabled Semantic Communication Systems," in *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663-2675, 2021

[4] A. Filali, A. Abouaomar, S. Cherkaoui, A. Kobbane and M. Guizani, "Multi-Access Edge Computing: A Survey," in *IEEE Access*, vol. 8, pp. 197017-197046, 2020

[5] W. Yang *et al.*, "Semantic Communication Meets Edge Intelligence," in *IEEE Wireless Communications*, vol. 29, no. 5, pp. 28-35, October 2022

[6] H. Zhang, H. Wang, Y. Li, K. Long and V. C. M. Leung, "Toward Intelligent Resource Allocation on Task-Oriented Semantic Communication," in *IEEE Wireless Communications*, vol. 30, no. 3, pp. 70-77, June 2023