

엣지 추론을 위한 은닉 차원 축소 기반 교통 혼잡 예측 모델 경량화

최민준, 김건민, 정세연, 김경백

전남대학교 인공지능융합학과

E-mail: larrychoi@jnu.ac.kr, geonminkim@jnu.ac.kr,

320eyeon@naver.com, kyungbaekkim@jnu.ac.kr

요 약

지능형 교통 시스템(Intelligent Transportation System, ITS) 환경에서 교통 혼잡 예측 모델은 높은 정확도뿐만 아니라 RSU(Roadside Unit), 신호제어기 등 제한된 장비에서 동작 가능한 경량성도 요구된다. 그러나 시공간 그래프 신경망을 단순히 축소하면 표현력이 부족해져 혼잡 이벤트 분류 성능이 저하될 수 있다. 본 논문은 기존 PIGNN-Event 구조를 기반으로, PINN(Physics-Informed Neural Network)으로 복원한 교통 밀도장 prior가 경량 GNN(Graph Neural Network)의 성능 저하를 완화할 수 있는지 분석한다. 실험 결과, 중간 경량 모델은 기존 모델과 거의 동일한 F1을 달성하면서 경량화에 성공하였다. 반면 최대 경량 모델에서는 성능이 크게 저하되어, 물리 prior가 경량화에 효과적이지만 최소한의 표현 용량은 필요함을 확인하였다.

I. 서 론

지능형 교통 시스템(Intelligent Transportation System, ITS)에서는 도로망 전역의 혼잡 상태를 실시간으로 예측하여 신호 제어, 경로 안내, 사고 대응 등을 지원한다. STGCN[1], DCRNN[2]과 같은 시공간 그래프 신경망은 도로망의 공간적 연결성과 시간적 변화를 함께 학습하여 교통 예측에 널리 활용되고 있으며, 최근 6G 기반 실시간 ITS 연구에서도 저지연 교통류 예측의 중요성이 강조되고 있다[3]. 그러나 실제 ITS 모델은 신호제어기, RSU 등 제한된 연산 자원을 가진 장비에서 동작해야 하므로, 예측 정확도뿐만 아니라 모델 크기와 추론 지연 시간도 중요한 요구사항이 된다.

기존에 진행한 PIGNN-Event 연구는 RSU와 Probe 차량이 혼재된 이기종 센서 환경에서, PINN으로 복원한 교통 밀도장을 물리 prior로 활용하여 3단계 혼잡 예측 성능을 향상시켰다[4]. 그러나 해당 연구는 모델 경량화 상황에서 물리 prior가 성능 저하를 얼마나 완화할 수 있는지는 충분히 분석하지 않았다. 한편 최근 물리 기반 모델은 물리 제약을 통해 순수 데이터 기반 모델이 학습해야 하는 구조적 패턴을 보완하고, 제한된 모델 용량에서도 물리적으로 일관된 표현을 제공할 수 있음을 보이고 있다[5, 6]. 특히 PINN 기반 prior는 교통류의 연속성과 혼잡 전파 구조를 명시적으로 반영하므로, 모델을 경량화하더라도 부족한 표현력

을 보완하여 성능 하락을 완화할 것으로 기대하였다.

따라서 본 논문은 기존 PIGNN-Event 구조를 기반으로, 경량 정도에 따른 네 가지 프로파일을 구성하고, PINN의 모델 용량 감소에 따른 성능 보전을 분석한다.

II. 본 론

1. 문제 설정

본 연구는 기존 선행 연구인 PIGNN-event를 기반으로, 모델 경량화가 교통 혼잡 예측 성능에 미치는 영향을 분석한다. 선행 연구는 PINN으로 복원한 교통 밀도장을 물리 prior로 사용하고(stage 1), 이를 STGCN 기반 분류기에 결합하여 5분 후 링크별 혼잡 단계를 0/1/2(원활/혼잡/심각)로 예측하는(stage 2) 구조이다. 본 논문에서는 물리 prior가 교통 혼잡 모델의 경량화에서 성능 보존에 기여하는지 검증한다. 따라서 표 1과 같이 기존 PIGNN-Event 설정인 full을 기준으로, Stage 1과 Stage 2의 은닉 차원을 단계적으로 축소한 세 가지 경량 프로파일을 구성하였다.

표 1. 경량화 프로파일 설정

	Stage1	Stage2
full	64	128
lite	48	64
lite_med	32	32
lite_max	16	16

2. 실험 환경 및 결과

표 2. 경량 성능 비교

	Params	MACs	Latency(ms)
STGCN	48,707	1.16B	25.23
full	205,542	4.61B	85.25
lite	59,158	1.22B	91.36
lite_med	18,534	0.34B	21.31
lite_max	5,734	0.10B	16.65

표 3. 제어 성능 비교

	F1	mIoU	Accuracy	pred1_wrong
STGCN	0.4700	0.3621	0.7552	0.4778
full	0.6021	0.4739	0.8011	0.3746
lite	0.6024	0.4738	0.8017	0.3686
lite_med	0.5901	0.4634	0.7978	0.3797
lite_max	0.3777	0.3053	0.7476	0.4326

실험은 광주 도로망 기반 시뮬레이션 환경에서 수행했다. 각 프로파일에 대해 3개 case와 10개 seed를 조합하여 30회 실험을 수행하였으며, 예측 성능은 accuracy, F1, mIoU, pred1_wrong으로 평가했다. pred1_wrong 지표는 모델이 성능을 위해 중간 단계

혼잡으로 과예측하는 것을 막기 위해 설정한 지표이다. 효율성은 파라미터 수, MACs, CPU forward latency를 기준으로 비교하였다.

표 2의 효율성 비교 결과, lite 계열 모델은 full 대비 파라미터 수, MACs, CPU forward latency를 전반적으로 줄였다. 특히 lite_med는 파라미터 수와 MACs를 각각 약 90.9%, 92.6% 절감하여, 제한된 ITS 엣지 장비에서의 추론 가능성을 높였다.

표 3에서 full은 STGCN 대비 높은 F1, mIoU, accuracy를 보여, 기존 연구와 같이 물리 prior가 혼잡 예측 성능 향상에 기여함을 확인하였다. 반면 lite_max는 F1이 크게 하락하였다. 이는 PINN이 경량화 과정의 표현력 손실을 완화할 수 있지만, 과도한 압축에서는 모델 판단 용량 부족으로 인한 붕괴를 의미한다.

III. 결 론

본 논문은 PINN 기반 교통 혼잡 예측 모델의 경량화에 있어 PINN이 성능 하락 방지에 어떠한 영향을 주는 지에 대해 분석하였다.

실험 결과, lite_med 모델은 full 대비 파라미터와 MACs를 크게 줄이면서도 성능을 유지하였다. 이는 물

리 정보가 경량 모델의 성능 보존에 효과적으로 작용할 수 있음을 보여준다. 반면 lite_max에서 성능이 크게 저하되어, 물리 prior가 있어도 최소한의 모델 용량은 필요함을 확인하였다.

감사의 글

본 연구성과는 2025년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(RS-2025-25398164)(33%)이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화 혁신인재양성사업의 지원을 받아 수행된 연구임(IITP-2026-RS-2022-00156287)(34%)본 연구는 2025년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 (S 3 4 7 1 3 9 2) (3 3 %)

References

- [1] J. Tang, L. Xia, and C. Huang, "Explainable spatio-temporal graph neural networks," Proc. 32nd ACM Int. Conf. on Information and Knowledge Management (CIKM 2023), 2023, pp. 2432-2441.
- [2] Y. Huang, et al., "Diffusion convolutional recurrent neural network with rank influence learning for traffic forecasting," Proc. 18th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications / 13th IEEE Int. Conf. on Big Data Science and Engineering (TrustCom/BigDataSE), 2019, pp. 678-685.
- [3] S. S. Chauhan et al., "6G conditioned spatiotemporal graph neural networks for real time traffic flow prediction," Scientific Reports, vol. 16, no. 1, 3902, 2026.
- [4] M. J. Choi, G. M. Kim, and K. B. Kim, "PIGNN-Event: 이기종 센서 환경에서 물리 기반 GNN을 이용한 교통 혼잡 예측," in Proc. of the 2026 KICS Summer Conference, 2026, 12C-P-55.
- [5] Z. Tao, H. Wang, and F. Liu, "LNN-PINN: A unified physics-only training framework with liquid residual blocks," Computer Physics Communications, 110237, 2026.
- [6] N. Kang et al., "PIG: Physics-informed gaussians as adaptive parametric mesh representations," in Proc. of the International Conference on Learning Representations (ICLR), 2025, pp. 73022-73046.