pm

2024년도 석사과정생연구장려금지원사업
신규과제 연구계획서

과제명	국문	신규 이질 간선 그래프 벤치마크에 기반한 그래프 딥러닝 연구
	영문	Research on graph neural networks based on a new heterophilic graph benchmark

◎ 작성 시 유의사항

- **"작성분량은 3P 이내"**로 작성하여야 하며, **위반할 경우 연구계획서 평가 시 초과분량(3P 초과)에 대한 평가 미실시 등 불이익을 받을 수 있음.**
  - ※ **작성분량은 1번~4번 항목에 대한 분량이며, 표지는 작성분량에서 제외**
- 내용작성과 관련한 설명내용**("작성요령(제출 시 삭제)")은 제거**하고 내용 기술
- 글자 크기에 대한 제한은 없으며, 가독성 등을 고려하여 연구책임자가 자유롭게 설정
- 파일에 DRM을 적용하거나 파일을 암호화 할 경우, 평가 시 불이익을 받을 수 있음

소속	**학부(과)** 컴퓨터과학 **전공** 컴퓨터과학	■ 석사과정 ▢ 석·박사 통합과정	1학기

1. 학업 계획

목표

• 그래프 데이터를 중심으로 한 복잡 구조 데이터에 대한 주요 최신 딥러닝 모델 및 성능 향상 기법에 대한 학업을 통하여 본 연구 과제를 성공적으로 실행하고 다양한 복잡 구조 데이터를 딥러닝으로 분석하는 능력을 배양하고자 함

학업 계획

기간	학업 내용	학업 성과
24년 1, 2월	• [Machine Learning with Graphs] 스탠포드 온라인 강의 수강 • [Machine Learning] 스탠포드 온라인 강의 수강 • [Introduction to Data Analytics] 코세라 온라인 강의 수강 • [Data Analysis and Visualization Foundations] 코세라 온라인 강의 수강	• 범용 머신러닝 기술과 그래프 데이터 학습 방법을 익힘 • 데이터 전처리 기술과 데이터 시각화 기술을 익히고 특히 시각화된 데이터를 자료를 효과적으로 전달하는 발표 기술을 익힘
24년 1학기	• [딥러닝입문] 대학원 강의 수강 • [자연어처리] 대학원 강의 수강 • [정보이론] 대학원 강의 수강	• 범용 딥러닝 기술을 익히고 대표적으로 딥러닝 기술이 이용되고 있는 자연어처리 분야의 연구 동향을 요약함 • 자연어처리 분야의 기술을 활용함으로써 그래프 데이터 학습의 성능을 향상시킬 방법을 고찰함 • 정보의 양적 개념을 익힘
24년 7, 8월	• 머신러닝/데이터 과학 커뮤니티인 캐글의 컴피티션 참여	• 실전 문제 해결 과정을 경험함
24년 2학기	• [그래프데이터분석] 대학원 강의 수강 • [클라우드컴퓨팅] 대학원 강의 수강	• 그래프 데이터 학습 분야의 최신 연구 동향을 요약함 • 빅데이터 분석을 위한 분산컴퓨팅 기술을 익힘

2. 연구 계획

연구 과제의 필요성

1) 가용 그래프 데이터의 종류와 양의 폭발적으로 증가와 그래프 딥러닝 기술의 발전

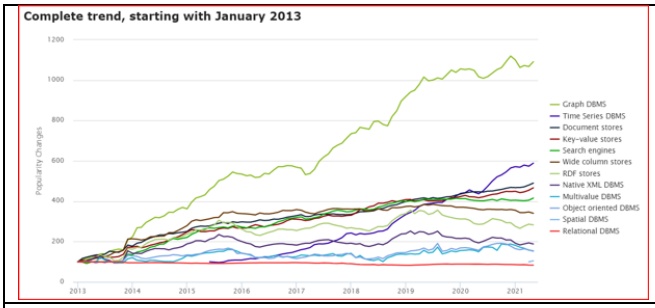

그림 2-1: 그래프 구조 데이터 수요 증가[DBEngine]

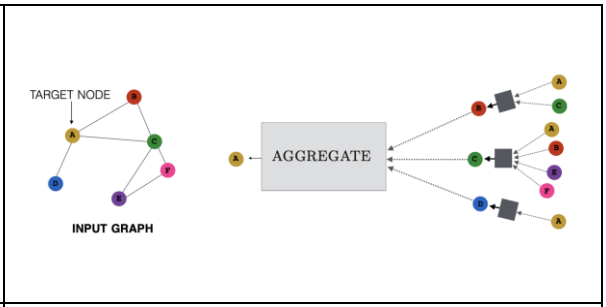

그림 2-2: 그래프 딥러닝의 핵심 기술인 메시지 패싱[GRL]

• 그래프 구조 데이터의 예시로는 추천 관련 구매 로그 데이터, 소셜 네트워크 데이터, 단백질 간 상호작용 데이터 등 다양하며 [그림 2-1]을 통해 알 수 있듯 최근 그래프 구조 데이터의 저장량이 빠르게 증가하고 있음

• 그래프 구조 데이터는 정점 자체의 특징 정보뿐 아니라 이웃 정점들과의 관계 또한 표현 가능함. 최근 [그림 2-2]와 같은 메시지 패싱 기반 그래프 딥러닝 기술을 통해 정점 특징과 정점의 컨텍스트를 동시에 고려한 고수준의 분석을 수행하고 있음. 그러나 기존 그래프 딥러닝은 동질성(Homophily, 연결된 정점은 같은 클래스에 속할 가능성이 높다)을 핵심 전제로 하므로 관련 없는 정점들의 임베딩까지도 과도하게 비슷해지는 문제(Over-smoothing)가 나타나기 쉬움

2) 이질 간선 그래프에 대한 기존 그래프 딥러닝의 성능 저하 현상과 이질 간선 그래프 벤치마크의 문제점[CLEH]

• **이질 간선 그래프에 대한 기존 그래프 딥러닝의 성능 저하 현상:** 이질 간선 그래프(Heterophilic Graph)는 [그림 2-3]과 같이 동질성에 반하는 간선, 즉 다른 클래스의 노드들을 연결한 간선들의 비율이 높은 그래프를 의미함. 기존 그래프 딥러닝은 동질성을 전제로 설계되어 있어 이질 간선 그래프에 대해 단순 다층 퍼셉트론보다 더 낮은 정확도를 보이기도 함을 실험을 통해 관찰함

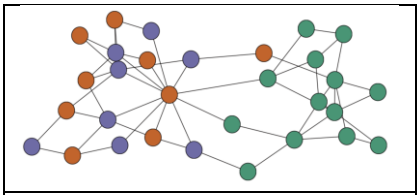

그림 2-3: 이질 간선 그래프 [CPGNN]

• **이질 간선 그래프 벤치마크의 문제점:** 최근 이질 간선 그래프에 대한 연구[ACM,

GloGNN, GREET]가 활발히 이루어지고 있음. 이들은 주로 6가지 이질 간선 그래프 데이터셋(WebKB 3개, Wikipedia 3개)을 벤치마크로 사용하여 성능을 측정하였음. 그러나 WebKB 데이터셋은 클래스 간 분포가 매우 불균형하였고 학습에 이용할 데이터의 양이 매우 적음. Wikipedia 데이터셋은 정점 특징과 구조 정보가 동일한 사본 정점 집합이 존재했으며 이로 인해 훈련 데이터와 테스트 데이터 사이의 정보 누수가 발생하였음[CLEH]. 누수 현상을 제거한 새로운 이질 간선 그래프 벤치마크(신규 이질 간선 그래프)를 이용하여 최신 이질 간선 그래프 딥러닝 모델의 성능을 재평가한 결과 Squirrel 데이터셋에서 최대 30%에 가까운 성능 하락이 발생함을 실험을 통해 직접 확인하였음

연구 과제의 목표 및 내용

1) 연구 과제의 최종 목표

• 신규 이질 간선 그래프 벤치마크에 기반하여 기존 그래프 딥러닝 모델들을 비교 분석하고 이를 바탕으로 (1) 지도 학습 환경과 (2) 비지도 학습 환경 각각에서 정점 분류 성능을 높이는 새로운 그래프 딥러닝 모델을 제안하고자 함

2) 연구 과제의 내용

[연구 주제 1] 신규 이질 간선 그래프 데이터에 대한 딥러닝 기반 지도 학습의 성능 개선

• 기존 이질 간선 그래프로부터 지도학습하는 경우, 최근 제안된 다중 채널에서 추출한 다양한 정보를 혼합하는 그래프 딥러닝 모델들[ACM, GloGNN, GPR]의 성능이 가장 우수한 것으로 알려져 있음

• 신규 이질 간선 그래프 데이터 상에서는 오히려 다중 채널 혼합 모델의 성능 저하가 나타남을 [그림 2-4]의 사전 실험 결과에서 확인함

• 다중 채널을 혼합하는 학습 방법은 이질 간선 그래프뿐 아니라 동질 간선 그래프에서도 우수한 성능을 보였음[ACM]. 따라서 동질 간선 그래프에 대해 다중 채널을 혼합하여 학습하는 방법이 우수한 이유를 우선적으로 분석하고자 함. 그리고 다중 채널을 이용해 기존 이질 간선 그래프를 학습하는 모델들의 장단점과 기존 그래프 데이터 상에서의 한계점을 조사하고 이를 바탕으로 신규 이질 간선 그래프에 대한 새로운 지도학습 모델을 설계하고자 함

• 연구 범위

① 동질 간선 그래프에 대한 다중 채널 혼합 학습의 우수성 분석

② 기존 이질 간선 그래프에 대해 다중 채널을 이용해 학습한 모델들의 장단점 비교 및 한계점 분석

③ 신규 이질 간선 그래프의 지도 학습 성능 개선을 위한 다중 채널 혼합 그래프 딥러닝 모델 설계

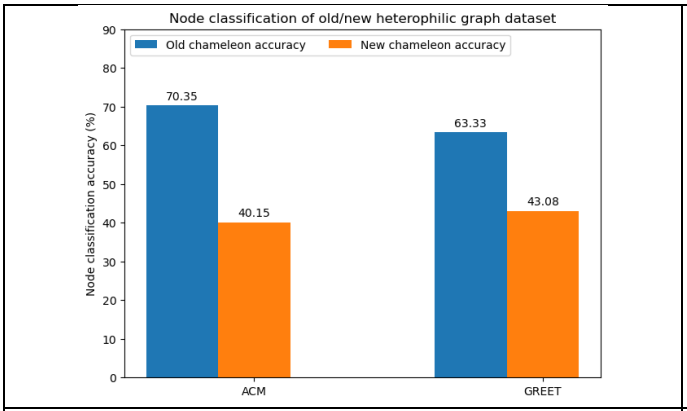

그림 2-4: 신규 이질 간선 그래프 데이터에 대한 성능 저하 확인 결과

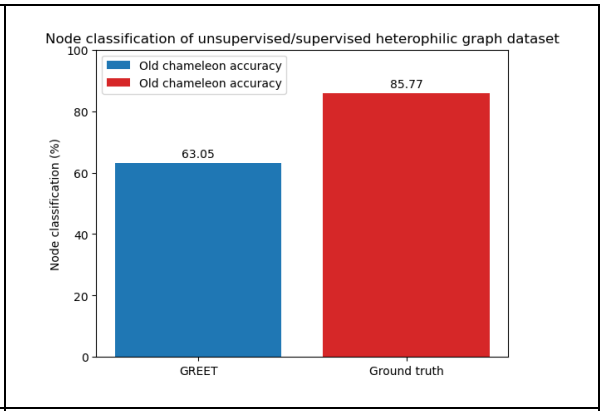

그림 2-5: 간선 성질 정보의 확실성에 따른 정점 분류 결과 비교

[연구 주제 2] 신규 이질 간선 그래프 데이터에 대한 딥러닝 기반 비지도 학습의 성능 개선

• 이질 간선 그래프를 비지도 학습하기 위해 간선의 이질성을 예측하는 모델을 부가적으로 학습하여 사용하고 있음[GREET]

• [그림 2-5]의 사전 실험 결과에 따르면 간선의 이질성 예측 정확도가 올라감에 정점 분류 성능이 큰 폭으로 향상되는 것을 확인할 수 있음

• 하지만 간선의 이질성을 예측하는 모델을 이용해 비지도 학습하는 모델 역시 신규 이질 간선 그래프에 대한 성능 저하가 발생함을 [그림 2-4]의 사전 실험 결과에서 확인함

• 대조 학습은 그래프 데이터 비지도 학습의 성능 향상을 위한 핵심적인 방법으로 주목받고 있음[GREET]. 따라서 정점에 대한 대조 학습 정확도와 간선의 이질성 예측 정확도 사이의 상관관계를 분석하고자 함. 그리고 클러스터링 기반 대조 학습 네트워크[CCGC]를 이용해 대조 학습에 이용할 샘플의 신뢰성을 높이고 이를 바탕으로 간선의 성질 예측 정확도를 향상시키고자 함

• 연구 범위

① 신규 이질 간선 그래프의 간선 성질 예측 정확도와 정점에 대한 대조 학습 사이의 상관관계 분석

② 클러스터링 기반 대조 학습 네트워크를 이용해 간선 성질 예측 정확도 개선 모델 설계

[연구 계획] 연간 연구 및 논문 발표 계획

	세부 연구 범위	상반기	하반기
연구 주제 1	동질 간선 그래프에 대한 다중 채널 혼합 학습의 우수성 분석	O	
	기존 이질 간선 그래프에 대해 다중 채널을 이용해 학습한 모델들의 장단점 비교 및 한계점 분석	O	
	신규 이질 간선 그래프의 지도 학습 성능 개선을 위한 그래프 딥러닝 모델 설계	O	
연구 주제 2	신규 이질 간선 그래프의 간선 성질 예측 정확도와 정점에 대한 대조 학습 사이의 상관관계 분석		O
	클러스터링 기반 대조 학습 네트워크를 이용한 간선 성질 예측 정확도 개선 모델 설계		O

• 상반기 연구 내용(연구 주제 1) 정리 및 피드백 수집을 위해 2024년 6월 '한국정보과학회 한국컴퓨터종합학술대회'에 논문을 발표하고자 함

• 연구 결과를 종합하여 2025년 'PAKDD' 학회(BK21 우수학회)에 논문 발표하는 것을 본 연구과제의 최종 목표로 함

3. 연구자 역량 및 연구 환경

연구자의 역량

분야	활동	세부 내용
문제 파악 및 해결 역량	한국정보과학회 학술대회 논문 2편 발표	• 학부연구생으로 수행한 그래프 딥러닝 관련 2가지 연구결과를 한국정보과학회 하계/동계 학술대회에 발표함 (논문 1) 「이질 클래스 간 링크와 신규 링크를 고려한 그래프 신경망 기반 추천 시스템」 (논문 2) 「간선 성질 예측과 다중 채널 혼합을 이용한 비지도 이질 간선 그래프 분석」
지식 습득 역량	컴퓨터과학부 전공과목 성적이 매우 우수함	• 군 전역 시점인 2학년 2학기 이후로 학기별 전공 평균 학점이 매우 높은 수준을 유지함 • 1-1(3.0), 1-2(3.75), 2-1(3.4), 2-2(3.9), 3-1(4.1), 3-2(4.125), 4-1(4.0)
연구 및 발표 역량	2023년 7월 한국정보과학회 하계 학술대회 우수발표논문상 수상	• 「이질 클래스 간 링크와 신규 링크를 고려한 그래프 신경망 기반 추천 시스템」가 우수발표논문으로 선정되어 확장본을 '정보과학회 컴퓨팅의 실제 논문지'에 투고한 상태임
	2024년 2월 한국정보과학회 동계 학술대회 우수발표논문상 수상	• 「간선 성질 예측과 다중 채널 혼합을 이용한 비지도 이질 간선 그래프 분석」이 우수발표논문으로 선정됨
	2023년 10월 교내 컴퓨터과학부 창작 작품 전시회 최우수상 수상	• 그래프 딥러닝 관련 프로젝트를 교내 창작 작품 전시회에 제출하여 최우수상을 수상함

연구 환경

그래프 딥러닝을 위한 고성능 하드웨어를 충분히 보유하고 있음	서울시립대 빅데이터센터 서버	Intel Xeon Gold 6348R 2.6GHz 56 cores / 1024GB MEMORY / SSD 480GB / Nvidia A10 4EA (36대) (해마다 신규 구매하여 확장하고 있음)
	소속 연구실 서버	GeForce 2080 8EA / 8GB MEMORY (2대)
교내 컴퓨터과학부 및 인공지능학부에 본 연구 관련 교수진 포진	컴퓨터과학부와 인공지능학부 내에 빅데이터분석, 음성인식, 인간중심인공지능(HCI), 컴퓨터비전, NLP, GPU 컴퓨팅 등 다양한 인공지능 관련 분야 연구 전문성을 보유한 18명의 교수진을 통하여 관련 대학원 교과목 수강을 통한 지도 및 직접적 연구조언이 가능한 환경임	

4. 기대 효과 및 진로 계획

연구 과제의 기대 효과	• 그래프 데이터의 성질에 독립적인 그래프 딥러닝 모델을 제시할 수 있음 • 레이블 정보에 대한 그래프 데이터 학습 의존성을 낮출 수 있음
진로 계획	• 목표: 그래프 구조 데이터 모델링부터 맞춤형 학습 방법까지 설계할 수 있는 그래프 딥러닝 전문가가 되고자 함 • 세부계획 - 석사 과정 중 이질 간선 그래프에 대한 지도/비지도 학습 성능 개선을 위한 딥러닝 모델을 설계함으로써 그래프 딥러닝에 대한 이해도를 높임 - 박사 과정 중 그래프 데이터의 성질을 고려한 추천 모델을 설계함으로써 그래프 데이터 분석의 실용적 사용 방안을 연구함