강좌상세정보

학점	개설학과	대표교수
3	데이터사이언스학전공	조요한(조교수) yohan.jo@snu.ac.kr

강좌정보

교과목번호	M3239.002300
교과목명(부제)	데이터사이언스 특강 (언어모델의 내부 작동 원리 - 해석 및 제어)
강좌번호	003
수업진행 언어	한국어
성적부여방식	A~F
성적평가방법 변경가능	NO

수업형태

교시별 수업형태 (강의실 동-호)	화(17:00~18:15) / 이론 / 43-411
	목(17:00~18:15) / 이론 / 43-411

파일 다운로드

첨부파일(국문)	[붙임2] 강의계획서 입력 양식(국문).pdf (47.84KB)
첨부파일(영문)	[붙임2] 강의계획서 입력 양식(영문).pdf (66.74KB)

강의 계획 상세

1. 수업목표	1. 언어모델의 내부 작동 원리에 대해 이해한다. 2. 언어모델의 내부 작동 원리를 해석할 수 있는 기술들을 이해하고 활용한다. 3. 언어모델의 내부 작동을 제어하는 기술들을 이해하고 활용한다.				
2. 교재 및 참고문헌					
3. 평가방법	성적부여방식	상대평가	등급제여부	A~F	
	출석(%)	0%			
	과제(%)	0%			
	중간(%)	0%			
	기말(%)	0%			
	수시평가(%)	0%			
	태도(%)	100%			

	기타(%)	0%				
	합계(%)	100%				
	출석 규정	수업일수의 1/3을 초과하여 결석하면 성적은 "F" 또는 "U"가 됨 (학칙 85조). 결석에 대하여 교원에게 별도로 출석인정을 받은 경우 예외로 할 수 있음 (학업성적처리규정. 조기취업자 출석 및 성적처리 지침).				
	기타사항					
4. 정원외신청	수용가능인원	최대 0 명				
5. 수강생 참고사항						
	면담시간/ 장소					
6. 강의계획	수업방식	☐ 플립러닝	☐ 이론위주 수업	☑ 토론위주 수업	☐ 프로젝트 수업	☐ 기타
		기타내용				

언어모델이 점점 강력해짐에 따라 연구자들은 언어모델이 학습한 표상을 해석하고 모델의 행동을 조정하여 보다 효과적으로 모델을 제어하고 신뢰성과 정렬성을 높이고자 한다. 본 교과목은 언어모델의 내부 메커니즘을 이해하고 수정하는 다양한 방법을 탐구한다. Vocabulary projection, sparse autoencoders, gradient-based methods와 같은 주요 해석 기법을 다루며, linear representation hypothesis, universal neurons, attention heads의 지식 저장 및 compositionality에서의 역할 등 중요한 연구 결과를 조명한다. 또한, 지식 편집과 벡터 연산과 같은 제어 기법을 공부하여 모델의 동작을 조정하는 방법을 배운다.

본 교과목은 학생 발표와 토론을 중심으로 진행된다. 매주 하나의 팀이 한 편의 논문을 두 번의 수업에 걸쳐 발표하며 토론을 주도한다. 발표 중에는 심층적인 질문과 토론을 통해 논문의 내용을 깊이 있게 탐구한다. 최종 성적은 발표의 우수성 및 토론 참여도에 의해 절대평가 방식으로 부여된다.

주제 (변동 가능)
1. Intepretation Methods for the Inner Workings of Language Models
- Vocabularity Projection
- Sparse Autoencoders
- Singular Value Deocmposition
- Gradient-Based Methods

2. Findings about the Inner Workings of Language Models
- Linear Representation Hypothesis
- Universal Neurons in Language Models
- Various Functions of Attention Heads
- Knowledge Storage and Retrieval
- Compositionality

3. Intervention Methods for the Inner Workings of Language Models
- Knowledge Editing
- Erasure
- Vector Operations

수강신청 시 참고사항
- 필요한 사전지식: 언어모델, 트랜스포머, 선형대수학, 기계학습 (본 수업에서 따로 다루지 않음)
- 수업 형식: 모든 수업은 학생들의 발표 및 심도있는 토론으로 진행됨.
- 추천 과목: 언어모델 및 대화형 인공지능에 대한 강의 및 실습을 원하는 학생들은 “대화형 인공지능” 과목을 추천.
- 성적: 절대평가
- 정원외 수강신청: 정원외 수강신청을 원하는 학생들은 정원외 수강신청 기간에 시스템에서 신청을 하고, 수강을 원하는 이유를 상세히 적을 것.

7. 장애학생 지원사항	강의 수강 관련	○ 시각장애: 교재 제작(디지털교재, 점자교재, 확대교재 등), 대필도우미 허용 ○ 지체장애: 교재 제작(디지털교재), 대필도우미 및 수업보조 도우미 허용 ○ 청각장애: 대필 및 문자통역 도우미 활동 허용, 강의 녹취 허용 ○ 건강장애: 질병 등으로 인한 결석에 대한 출석 인정, 대필도우미 허용 ○ 학습장애: 대필도우미 허용 ○ 지적장애/자폐성장애: 대필도우미 및 수업 멘토 허용

	과제 및 평가 관련	◯ 시각장애/지체장애/청각장애/건강장애/학습장애: 과제 제출기한 연장, 과제 제출 및 응답 방식의 조정, 평가 시간 연장, 평가 문항 제시 및 응답 방식의 조정, 별도 고사실 제공 ◯ 지적장애/자폐성장애: 개별화 과제 제출 및 대체 평가 실시
	비고	본 강의를 수강하는 장애학생들에게는 이상의 지원 서비스 이외에도 장애학생 개개인의 특성과 요구에 따라, 지도교수 및 장애학생지원센터와의 상담을 통하여 적절한 수준의 지원 서비스를 제공합니다. 장애학생에 대한 지원서비스와 관련하여 문의사항이 있는 학생들은 담당교수 혹은 장애학생지원센터(02-880-8787)로 문의바랍니다.

주차별 강의계획

주차구분	주차별 강의계획 내용
주차별 강의계획 내역이 없습니다.	