

4190.773: Multimodal Deep Learning Theories and Applications

Sangdoo Yun^{1,2} Jin-Hwa Kim^{1,2} Jiyoung Lee¹
¹NAVER AI Lab ²Seoul National University

Fall 2024

Overview

This course is designed for graduate students interested in developing theories and applications of multimodal deep learning. The course will cover core concepts and fundamental theories that are applied in various enterprise applications. Assuming prior knowledge in machine learning and related topics, the course will explore multimodal deep learning from vision and language to audio and speech using deep neural networks and the techniques used in deep learning literature. If time permits, an emerging topic of neural graphics will also be discussed, focusing on studies on the NeRFs in an aspect of multi-view representation learning.

1 Target Audience

- Graduate students with prior knowledge in machine learning and related courses who are looking to specialize or broaden in multimodal deep learning.

2 Course Objectives

- Gain a deep understanding of multimodal deep learning theories and applications
- Explore various modalities such as vision, language, audio, and speech in deep learning
- Stay updated with emerging topics in neural graphics, specifically NeRFs

3 General Information

3.1 Class information

- Friday 1:00 PM - 3:50 PM, Sep. 6th to Dec. 13th
- Bldg. 302, Room 107
- Credit 3-3-0

3.2 Office hour

- Friday 10:00 AM, AIIS Bldg. 942, Room 208
- Reservation only

– Sangdoo Yun (sangdoo.yun@navercorp.com, Jin-Hwa Kim (jnhwkim@snu.ac.kr), Jiyoung Lee (lee.j@navercorp.com)

4 Grade

- Quiz (30%)
 - Quizzes are regularly administered at the start of each class, with intervals of one to three weeks based on class progress and project sessions.
 - Attendance is crucial, as missing over one-third of the quizzes without a valid excuse may fail.
 - The final score is calculated by averaging the scores of all quizzes taken, excluding the lowest and highest scores.
 - Quizzes may be considered in place of the attendance score.
- Project (proposal 30% + final 40% = 70%)
 - Develop a new research idea for multimodal deep learning.
 - Each team consists of *two* to *three* collaborators to prepare two (proposal and final) presentations.
 - Presentation must include the following contents: problem statement, data collection or preparation, model architecture, and theoretical and experimental analysis.
 - Rubrics for the assessments will be notified before the presentation dates.
 - Recording is only permitted for conference or workshop presenters.

5 Schedule

The following is a tentative schedule for the course:

Week	Date	Topics
1	6 Sep	Introduction
2	13 Sep	Multimodal tasks and datasets
3	20 Sep	Multimodal representation learning
4	27 Sep	Multimodal reasoning and interpretability
5	4 Oct	Multimodal pre-training 1
6	11 Oct	Multimodal pre-training 2
7	18 Oct	Project proposal
8	25 Oct	Multimodal generation 1
9	1 Nov	Multimodal generation 2
10	8 Nov	Topics in audio and video
11	15 Nov	Neural graphics
12	17 Nov	Topics in neural graphics
13	22 Nov	Topics in trustworthy AI
14	29 Nov	Final presentations
15	6 Dec	Final presentations
16	13 Dec	No class

Note: The course content and sequence may be adjusted based on the duration of the lecture and the interests of the students. Additional readings, assignments, and projects may also be included to enhance the learning experience.