科目ナンバリング	U-LAS30 20046 LE10						
授業科目名<英訳>	Multimodal AI: Unifying Vision, Language and Audio-E2 Multimodal AI: Unifying Vision, Language and Audio-E2			担当者所属職名・氏名	情報学研究科 特定准教授 LALA, Divesh Kanu		
群	情報学科目群			分野(分類)	(各論)	使用言語	英語
旧群	B群	単位数	2単位	週コマ数	1コマ	授業形態	講義（対面授業科目）
開講年度・開講期	2025・後期	曜時限	木3	配当学年	全回生	対象学生	全学向

[授業の概要・目的]

The development of powerful models such as ChatGPT and speech recognition has meant AI now exhibits more human-like intelligence. However, machines also need to take into account all of the human senses including sight, sound and even touch. In this course, students will gain an understanding of important AI models currently being used in the fields of vision, speech and language. We then discuss how we can take two or more of these fields and combine them to create multimodal models. There will also be the opportunity to practically test these AI models and understand where they work and what needs to be improved.

[到達目標]

Students will gain an understanding of machine learning techniques and architectures and also how to construct basic models. They will also know how to properly evaluate AI models using real data.

[授業計画と内容]

1. Introduction to modelling (1 week)
Recent AI models have been responsible for many technological breakthroughs. How does a machine use a model to extract data and process it intelligently? We discuss how AI models possess human capabilities to process and understand real-world data.

2. Overview of computer vision, language and speech (3 weeks)
We review the historical techniques and important models in the domains of vision, speech and language. Students should understand the fundamental concepts behind the models and what the current challenges are in these fields.

3. Introduction to modelling (2 weeks)
Basic construction of AI models will be discussed, with the focus on neural networks as they are responsible for much of the state-of-the-art AI which is used today. Students will be able to see some of these models in action.

4. Transformer models (2 weeks)
AI has been largely improved through the concept of transformer models. An overview of this architecture will be provided to understand what makes it such a powerful technique that has been adapted to many AI applications.

5. Multimodal models and techniques (3 weeks)
In these lectures introduce the concept of combining different data streams, known as multimodality. What situations do we need it and why can it further improve AI models? We take a look at techniques for creating multimodal models which can combine two or more of vision, speech and language. Topics include

Multimodal AI: Unifying Vision, Language and Audio-E2(2)へ続く↓↓↓

multimodal representation and fusion.

6. Multimodal interaction (3 weeks)
Interaction between humans and computers or robots will be discussed as multimodal interaction can greatly improve the interaction experience. User interfaces and human-agent conversation will be emphasized. Students will also have the opportunity to create their own models or system and evaluate the results.

7. Feedback (1 week)

[履修要件]

特になし

[成績評価の方法・観点]

Attendance and participation (30%), exercises (50%) and a final report (20%).

[教科書]

Handouts

[授業外学修（予習・復習）等]

Students should aim to review course content for 30 minutes before and after class and try to read fundamental papers which will be presented during classes.

[その他（オフィスアワー等）]