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	2026・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 multiple types of data, known as modalities. In this course, students will gain an understanding of important AI models currently being used in the fields of vision, language and speech. 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. Accompanying the lectures in this course will be code in Python so students can try implementing these models for themselves.

[到達目標]

Students will gain a broad understanding of state-of-the art multimodal models and the techniques which are used to create intelligent systems. They will also learn to approach problems multimodally to improve model performance.

[授業計画と内容]

1. Introduction to multimodality (1 week)
We introduce the concept of modalities. What is a modality? How are modalities used in modern AI? We introduce some common multimodal tasks and describe how the course will teach students how AI can achieve these tasks.

2. Vision (3 weeks)
Students will first learn the basic modeling process, using image recognition as an example. We then introduce a foundational machine learning model, the neural network, and explain why this has become the basis for almost all modern AI. We show how the field of computer vision has built on this model through the convolutional neural network (CNN).

3. Language and audio (2 weeks)
Language and audio are two other modalities which have made a huge impact on modern AI. In these lectures we show how the neural network can be adapted to accommodate these modalities by introducing the recurrent neural network (RNN) and related models.

4. Transformers and attention (1 week)
Transformers are the foundation of large language models such as ChatGPT. In this lecture we will describe the important concept of attention and the transformer architecture which has revolutionized AI.

5. Advanced architectures and techniques (1 week)
In this lecture we take a special look at how neural network-based architectures can be extended to accomplish many different tasks through the use of techniques such as fine-tuning. We will also discuss latent

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

spaces, which are necessary for understanding multimodal models.

6. Multimodal techniques (4 weeks)
We now look at models which combine different modalities, known as multimodal models. For these lectures we take an in-depth look at topics related to these multimodal models and take a specific focus on combining vision and language. These architectures have allowed modern AI to achieve tasks such as describing and answering questions about an image. Generative AI models will be introduced during these lectures. Students will also have the opportunity to try out multimodal LLMs on their own computers.

7. Other applications (2 weeks)
Applications related to multimodality will be described, including conversation and multimodal interfaces.

8. Final exam

9. Feedback

[履修要件]

特になし

[成績評価の方法・観点]

Grades will be equally split (33% each) between attendance and participation, an assignment and a final exam.

[教科書]

未定

[授業外学修（予習・復習）等]

Students should aim to review course content through resources and code provided during the course on LMS.

[その他（オフィスアワー等）]

[主要授業科目（学部・学科名）]