CMU ONLINE
Carnegie Mellon University

Syllabus Overview Sample: Multimodal Machine Learning

Program: Generative AI and Large Language Models

Faculty: Yonatan Bisk, Assistant Professor, Language Technologies Institute

David Mortensen, Assistant Research Professor, Language Technologies Institute

Ralf Brown, Principal Systems Scientist, Language Technologies Institute

Course Number: 11977 **Units:** 12

Zoom Class: Weekday Evening, ET

Course Description

Multimodal machine learning (MMML) is a vibrant interdisciplinary research area that advances core goals of artificial intelligence by integrating and modeling multiple communicative modalities, including language, speech, vision, and other sensory signals. Originating from early work in audio-visual speech recognition and evolving through modern language-and-vision systems such as image and video captioning, MMML addresses unique challenges arising from heterogeneous data sources and complex interdependencies between modalities.

This course introduces the fundamental mathematical and algorithmic concepts in machine learning and deep learning that underpin the **six central challenges of multimodal machine learning**:

1. Representation
2. Alignment
3. Reasoning
4. Generation
5. Transference
6. Quantification

Students will study both foundational and modern approaches to these challenges, including multimodal autoencoders, deep canonical correlation analysis, multimodal tensor fusion, multi-kernel learning, attention mechanisms, multimodal recurrent models, multimodal transformers, neuro-symbolic methods, mutual information–based objectives, and multimodal graph neural networks.

The course also examines state-of-the-art probabilistic models and computational algorithms through readings of recent research literature, with attention to open problems and emerging directions in the field. Applications discussed throughout the course include multimodal affect

recognition, language grounding, language-vision navigation, image and video captioning, and cross-modal multimedia retrieval.

Course Learning Objectives

By the end of this course, students will be able to:

- Analyze multimodal data (language, audio, vision) to identify challenges in representation, alignment, reasoning, and generation.
- Apply and evaluate mathematical foundations and state-of-the-art models (e.g., multimodal transformers, fusion techniques, graph networks) to multimodal machine learning tasks.
- Critique and design multimodal AI approaches by assessing research findings, addressing limitations, and proposing innovative applications.

Faculty Introduction



Yonatan Bisk is an Assistant Professor at Carnegie Mellon University's Language Technologies Institute (LTI) and Robotics Institute. His research focuses on the intersection of natural language processing (NLP), robotics, and embodied AI, with an emphasis on enabling machines to interpret and act upon language in the context of physical and social environments. Dr. Bisk's work advances theory of mind and embodied language systems, exploring how language interacts with vision, action, and reasoning.



David R. Mortensen is an Assistant Research Professor in the Language Technologies Institute at Carnegie Mellon University. He received his PhD in Linguistics from the University of California at Berkeley and held a position as an Assistant Professor of Linguistics at the University of Pittsburgh before transitioning to CMU, computational linguistics, and NLP. His work sits at the intersection between natural language processing, speech processing, phonology, morphology, and language change. He has published widely in both linguistics and CL/NLP venues and received a Best Paper Award at ACL 2025.



Dr. Ralf Brown is a Principal Systems Scientist at Carnegie Mellon University's Language Technologies Institute (LTI). With a Ph.D. in Computer Science from Carnegie Mellon, Dr. Brown has been a pivotal member of the institute since 1993, contributing to groundbreaking advancements in machine translation, language identification, and computational linguistics. His distinguished career

has been recognized with awards such as the IJCAI Distinguished Paper Award and Carnegie Mellon's Allen Newell Award for Research Excellence.

Course Topics

Topic	Learning Objectives
Introduction to Multimodal Machine Learning	• Define "modality" and "multimodal" in both perceptual and computational contexts. • Describe different types of human and machine perception signals. • Explain the challenges of combining and interpreting heterogeneous data sources. • Identify the six core challenges in multimodal machine learning. • Explain course logistics, expectations, and project structure.
Unimodal Representations	• Explain the motivation for learned representations across language, vision, audio, and sensor data. • Compare unimodal representation methods and architectures, including embeddings, convolutional models, and sequence models. • Analyze design trade-offs in unimodal models related to efficiency, data, and semantic content. • Assess the assumptions, biases, and transferability of representations. • Critically evaluate model interpretability techniques, relating visualization and performance metrics to real-world applicability and domain relevance.
Multimodal Fusion	• Evaluate conceptual frameworks and fusion architectures. • Apply algebraic principles to solve computational challenges. • Implement and analyze dynamic and adaptive fusion control. • Diagnose bias and design for stability in multimodal systems.

Alignment	● Explain how multimodal models identify and represent relationships across modalities through discrete, continuous, and contextual alignment methods. ● Apply core alignment techniques, such as attention mechanisms, dynamic time warping, and discretization, to analyze how models synchronize multimodal data. ● Analyze how alignment and contextual representation principles serve as conceptual foundations for transformer-based multimodal models.
Transformers	● Explain how transformer architectures internalize alignment, interaction, and loss design in multimodal learning. ● Analyze design tradeoffs in multimodal transformers, including when and how to restrict cross-modal attention. ● Compare vision transformers and convolutional models by reasoning about inductive bias, patch size, and compute tradeoffs. ● Evaluate the implications of discretization and generation in vision transformers. ● Critically assess the assumptions underlying language-first versus vision-first multimodal transformer designs.
Generation	● Explain generative modeling as learning and sampling from probability distributions, including how conditioning, priors, and underspecification shape generated outputs across modalities. ● Analyze how representational choices affect generation, such as language-bound representations, latent variables, tokenization, and diffusion processes, and why these choices lead to different strengths and failure modes. ● Diagnose common failures in generative systems, distinguishing between model limitations and task or evaluation design issues (e.g., underspecification, compositional errors), and reason about why scaling alone does not resolve these challenges.
Video Transformers + Dialogue +	● Analyze how multimodal representations are learned from weakly paired video–language data, and explain

Reasoning	how alignment constraints, hard negatives, and abstraction shape what becomes semantically meaningful. ● Evaluate how architectural choices for modeling time (e.g., 2D vs 3D convolution, separable designs, bottom-heavy vs top-heavy modeling) reflect assumptions about where information lives and how much computation is required to capture it. ● Compare neural, symbolic, and hybrid approaches to reasoning and interaction, and assess how imposing structure changes interpretability, flexibility, and brittleness in dialogue and neurosymbolic systems.
Modality Biases	● Explain how distributional skew in training data can reshape a model's decision boundary, leading to systematic errors even when the relevant modality contains clear evidence. ● Analyze why rare counterexamples (counterfactuals) exert limited influence under standard training, and contrast this with how humans update their internal models. ● Evaluate the tradeoff between robustness and performance, and explain why adding noise, more data, or additional modalities does not automatically eliminate modality bias.
Multimodal LLM Agents	● Articulate what new capabilities LLM agents enable, particularly for tasks requiring planning, tool use, and interaction with environments. ● Identify the key technical challenges of building agent systems, especially in real-world multimodal settings. ● Explain approaches researchers use to address these challenges, including methods for evaluation, planning, memory, and reusable skills.
New Audio Research Directions in MMML	● Explain how speech encodes multiple layers of information beyond words, including prosody, timing, and spectral structure. ● Describe how speech interacts with other modalities, such as facial expressions and gestures, in both perception and production. ● Discuss how multimodal integration enables new research directions in audio-based systems,

	particularly for tasks involving emotion, interaction, and human-centered communication.
Embodiment | Navigation	● Explain why embodiment is necessary for acquiring grounded semantics (i.e., how interaction enables richer understanding than passive observation – embodied grounding). ● Evaluate algorithmic strategies used in continuous spatial navigation, specifically focusing on attention-alignment (progress monitors), error recovery (backtracking), and techniques used to transfer a learned policy from simulation to real world. ● Justify the integration of non-visual sensory modalities (such as force and tactile feedback) to overcome underspecification of human language in physical manipulation tasks. ● Compare methods of action representation, such as predicting 3D voxels in empty space versus tokenizing continuous robot movements. ● Evaluate the strengths and inherent ambiguities of different modalities used to represent goal states (e.g., language, video demonstrations, audio, and sketches), recognizing that specifying an abstract goal requires overcoming severe underspecification and inferring human intent.
Final Presentations & Report	● Analyze multimodal data (language, audio, vision) to identify challenges in representation, alignment, reasoning, and generation. ● Apply and evaluate mathematical foundations and state-of-the-art models (e.g., multimodal transformers, fusion techniques, graph networks) to multimodal machine learning tasks. ● Critique and design multimodal AI approaches by assessing research findings, addressing limitations, and proposing innovative applications.

NOTE: This is a sample designed to provide an overview of topics covered in the course based on previous course syllabi. All students will be provided an updated and final syllabus at the start of the course. Faculty, content, and schedule are all subject to change.