Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

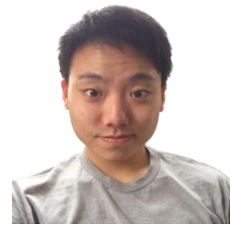
Lecture 1.1: Introduction

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Your Instructor and TAs This Semester (11-777)

Louis-Philippe Morency
morency@cs.cmu.edu
Course lecturer



Paul Liang
pliang@andrew.cmu.edu
TA & guest lecturer

Prakhar Gupta
prakharg@cmu.edu
TA

Martin Q. Ma
qianlim@cmu.edu
TA

Shikib Mehri
amehri@andrew.cmu.edu
TA

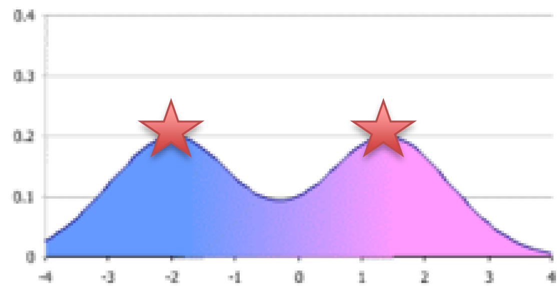
Torsten Wörtwein
twoertwe@cs.cmu.edu
TA

Lecture Objectives

- Introductions
- What is Multimodal?
  - Multimodal communicative behaviors
- A historical view of multimodal research
- Core technical challenges
  - Representation, translation, alignment, fusion and alignment
- Course syllabus and project assignments
  - Grades and course structure

What is Multimodal?

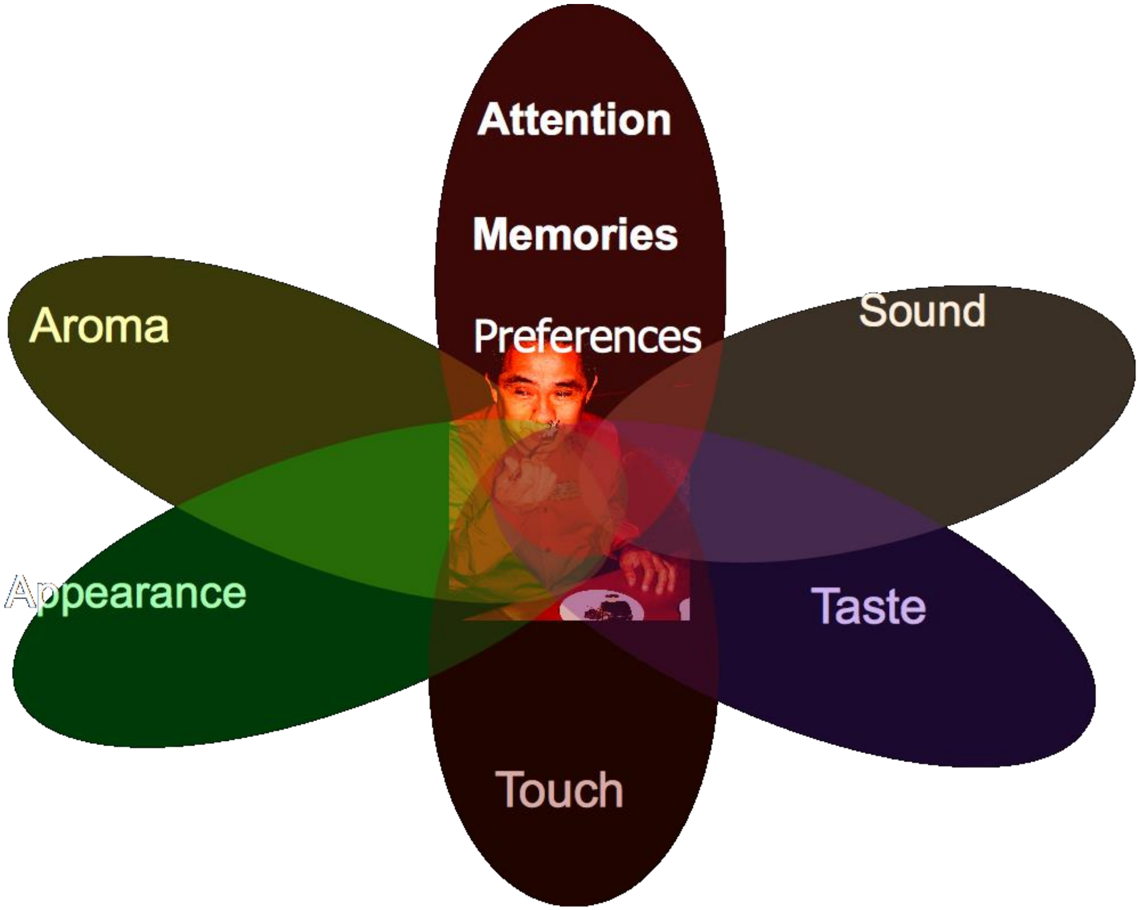
What is Multimodal?



Multimodal distribution

- Multiple modes, i.e., distinct “peaks” (local maxima) in the probability density function

What is Multimodal?



Sensory Modalities

Multimodal Communicative Behaviors

Verbal

Lexicon
- Words

Syntax
- Part-of-speech
- Dependencies

Pragmatics
- Discourse acts

Vocal

Prosody
- Intonation
- Voice quality

Vocal expressions
- Laughter, moans

Visual

Gestures
- Head gestures
- Eye gestures
- Arm gestures

Body language
- Body posture
- Proxemics

Eye contact
- Head gaze
- Eye gaze

Facial expressions
- FACS action units
- Smile, frowning

What is Multimodal?

Modality

The way in which something happens or is experienced.

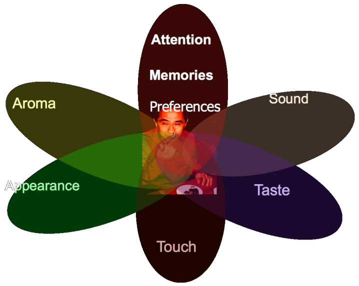
- *Modality* refers to a certain type of information and/or the representation format in which information is stored.
- *Sensory modality:* one of the primary forms of sensation, as vision or touch; channel of communication.

Medium ("middle")

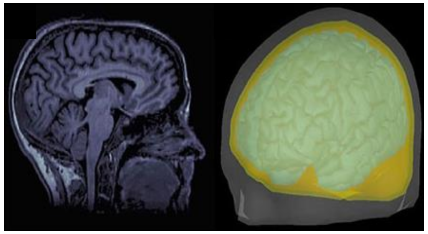
A means or instrumentality for storing or communicating information; system of communication/transmission.

- *Medium* is the means whereby this information is delivered to the senses of the interpreter.

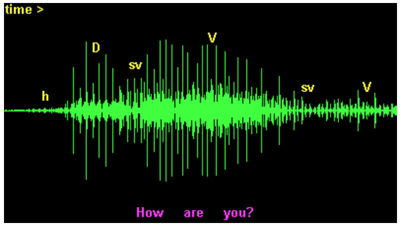
Multiple Communities and Modalities



Psychology



Medical

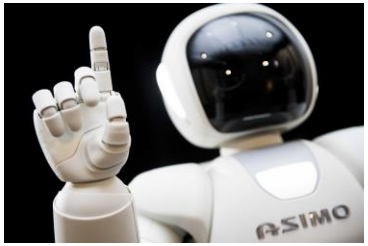


Speech

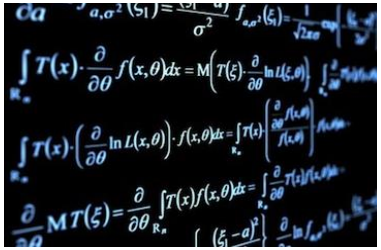
Vision

Language

Multimedia



Robotics



Learning

Examples of Modalities

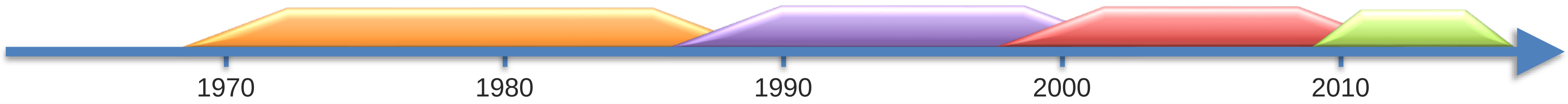
- Natural language (both spoken or written)
- Visual (from images or videos)
- Auditory (including voice, sounds and music)
- Haptics / touch
- Smell, taste and self-motion
- Physiological signals
  - Electrocardiogram (ECG), skin conductance
- Other modalities
  - Infrared images, depth images, fMRI

A Historical View

Prior Research on "Multimodal"

Four eras of multimodal research

- The "behavioral" era (1970s until late 1980s)
- The "computational" era (late 1980s until 2000)
- The "interaction" era (2000 - 2010)
- The "deep learning" era (2010s until …)
  - Main focus of this course

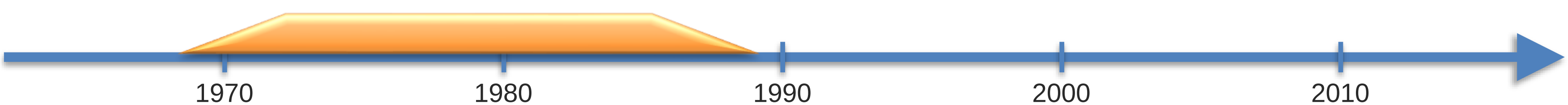


Language and Gestures

David McNeill
University of Chicago
Center for Gesture and Speech Research

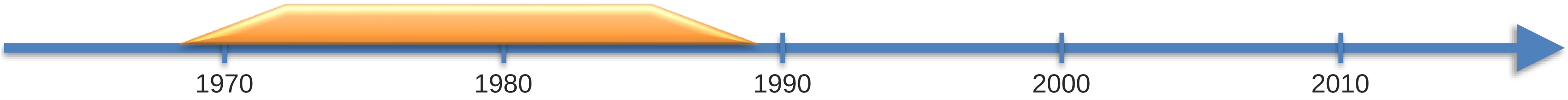
“For McNeill, gestures are in effect the speaker’s thought in action, and integral components of speech, not merely accompaniments or additions.”

- TRIVIA: Justine Cassell was a student of David McNeill



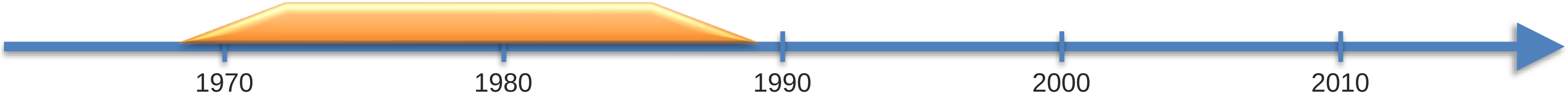
The McGurk Effect (1976)

[Hearing lips and seeing voices – Nature]



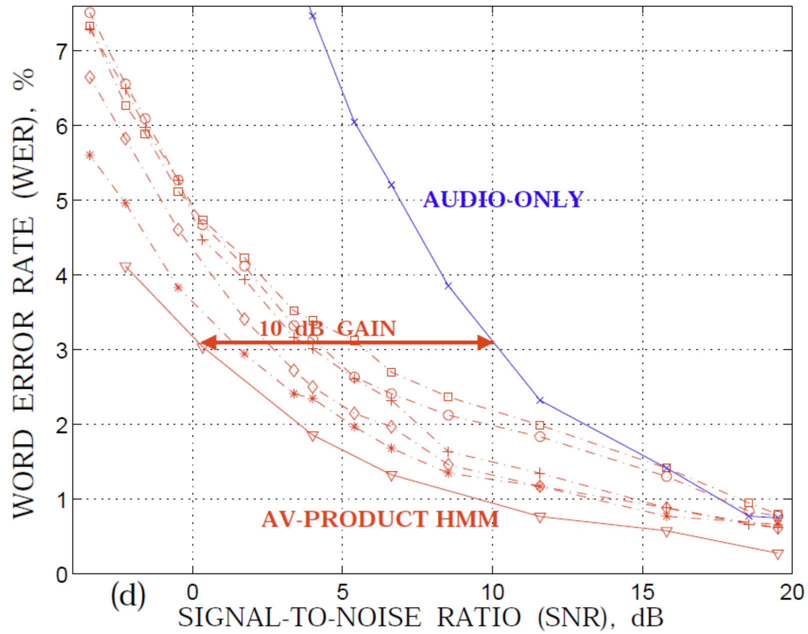
The McGurk Effect (1976)

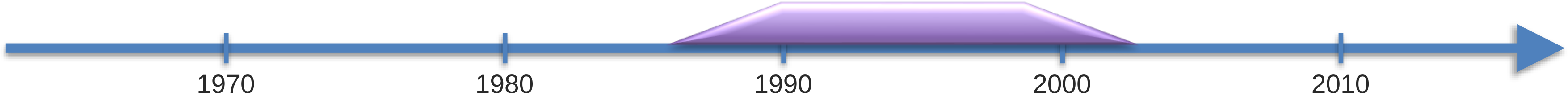
[Hearing lips and seeing voices – Nature]



➢ The "Computational" Era(Late 1980s until 2000)

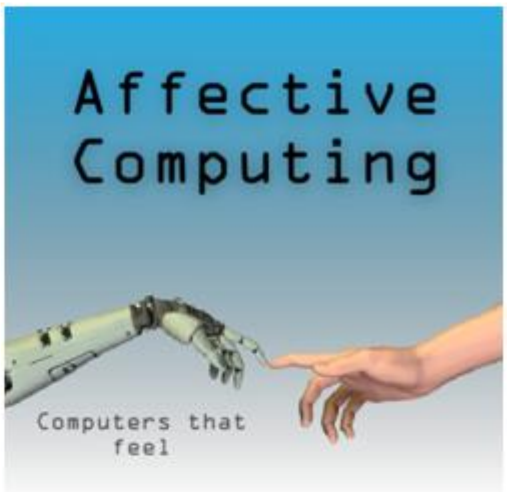
1) Audio-Visual Speech Recognition (AVSR)





➢ The “Computational” Era (Late 1980s until 2000)

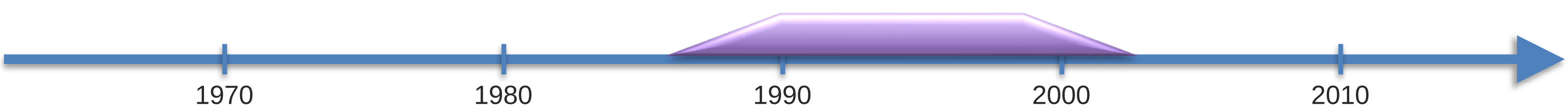
2) Multimodal/multisensory interfaces



Rosalind Picard

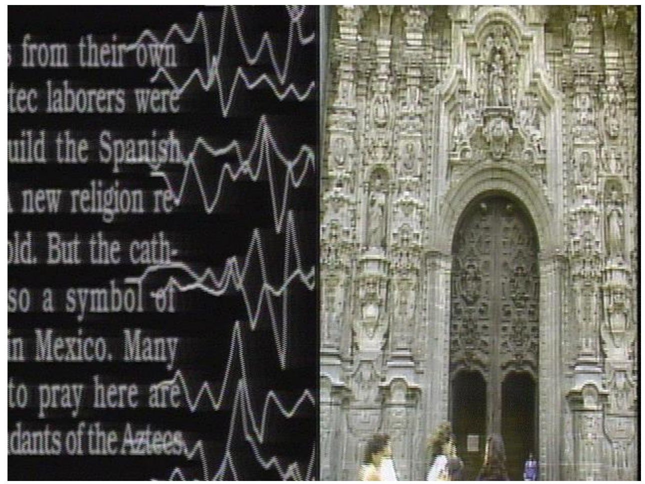
Affective Computing *is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena.*

❑ TRIVIA: Rosalind Picard came from the same group (MIT, Sandy Pentland)



➢ The "Computational" Era (Late 1980s until 2000)

3) Multimedia Computing

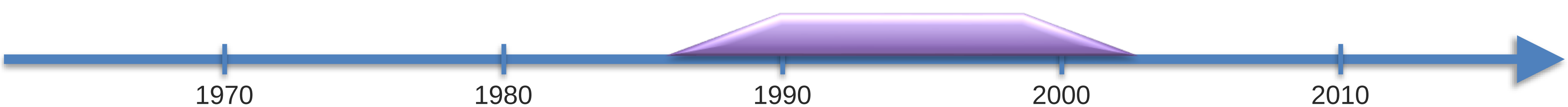


Carnegie Mellon University

informedia
digital video understanding

[1994-2010]

"The Informedia Digital Video Library Project automatically combines speech, image and natural language understanding to create a full-content searchable digital video library."



➢ The "Interaction" Era (2000s)

1) Modeling Human Multimodal Interaction

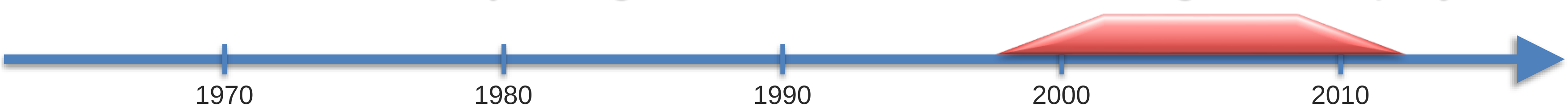
AMI Project [2001-2006, IDIAP]

- 100+ hours of meeting recordings
- Fully synchronized audio-video
- Transcribed and annotated

CHIL Project [Alex Waibel]

- Computers in the Human Interaction Loop
- Multi-sensor multimodal processing
- Face-to-face interactions

❑ TRIVIA: Samy Bengio started at IDIAP working on AMI project



➢ The “Interaction” Era (2000s)

1) Modeling Human Multimodal Interaction

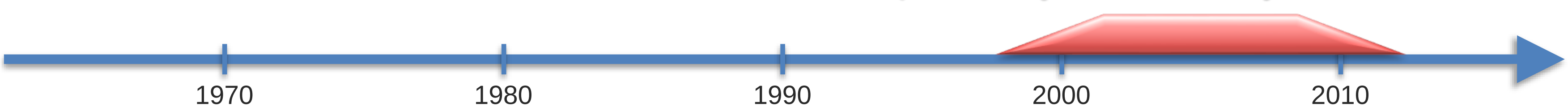
CALO Project [2003-2008, SRI]

- Cognitive Assistant that Learns and Organizes
- Personalized Assistant that Learns (PAL)
- Siri was a spinoff from this project

SSP Project [2008-2011, IDIAP]

- Social Signal Processing
- First coined by Sandy Pentland in 2007
- Great dataset repository: http://sspnet.eu/

❑ TRIVIA: LP’s PhD research was partially funded by CALO ☺



➢ The “deep learning” era (2010s until …)

Representation learning (a.k.a. deep learning)

- Multimodal deep learning [ICML 2011]
- Multimodal Learning with Deep Boltzmann Machines [NIPS 2012]
- Visual attention: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [ICML 2015]

Key enablers for multimodal research:

- New large-scale multimodal datasets
- Faster computer and GPUS
- High-level visual features
- “Dimensional” linguistic features

Our course focuses on this era!

Core Technical Challenges

Core Challenges in “Deep” Multimodal ML

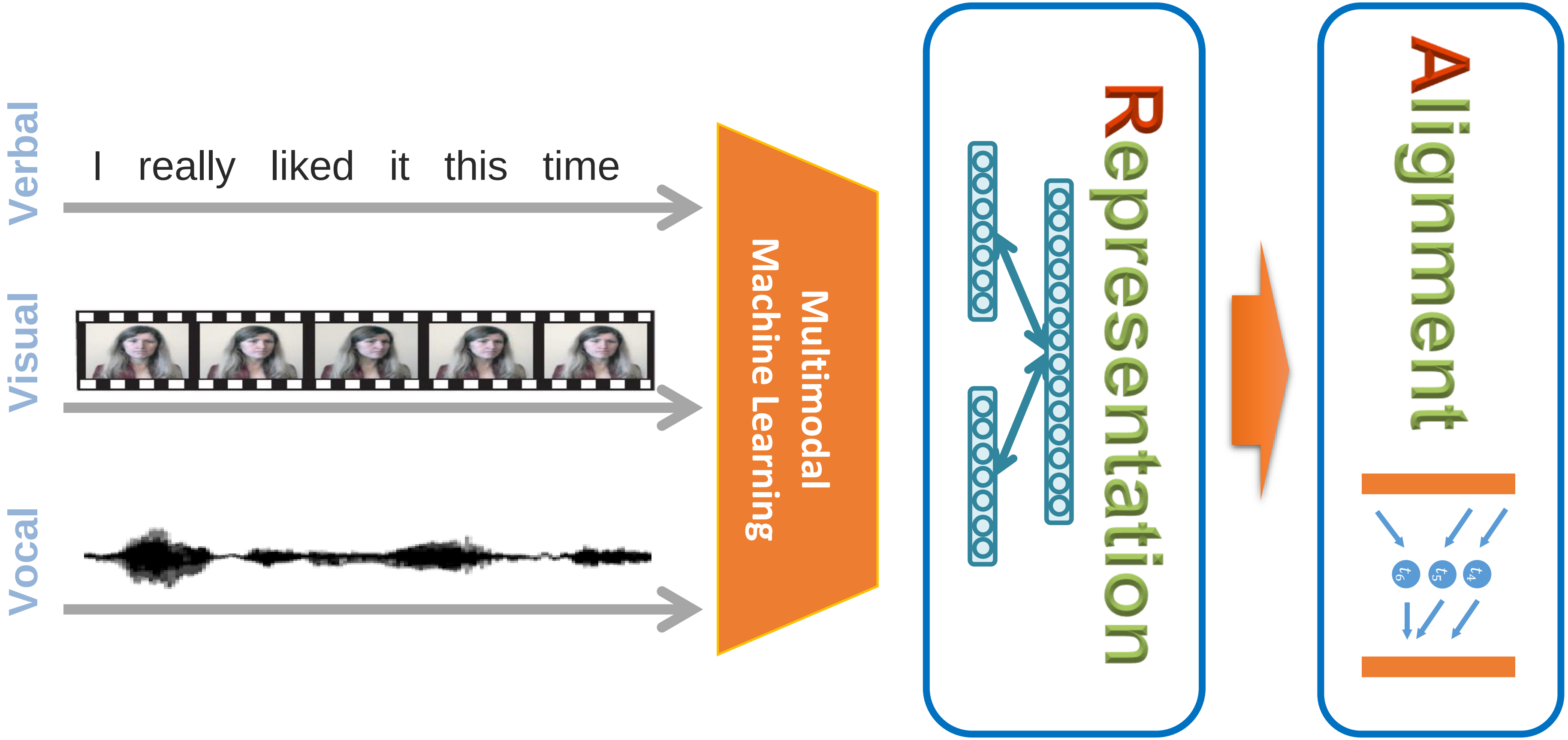
Multimodal Machine Learning: A Survey and Taxonomy

By Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

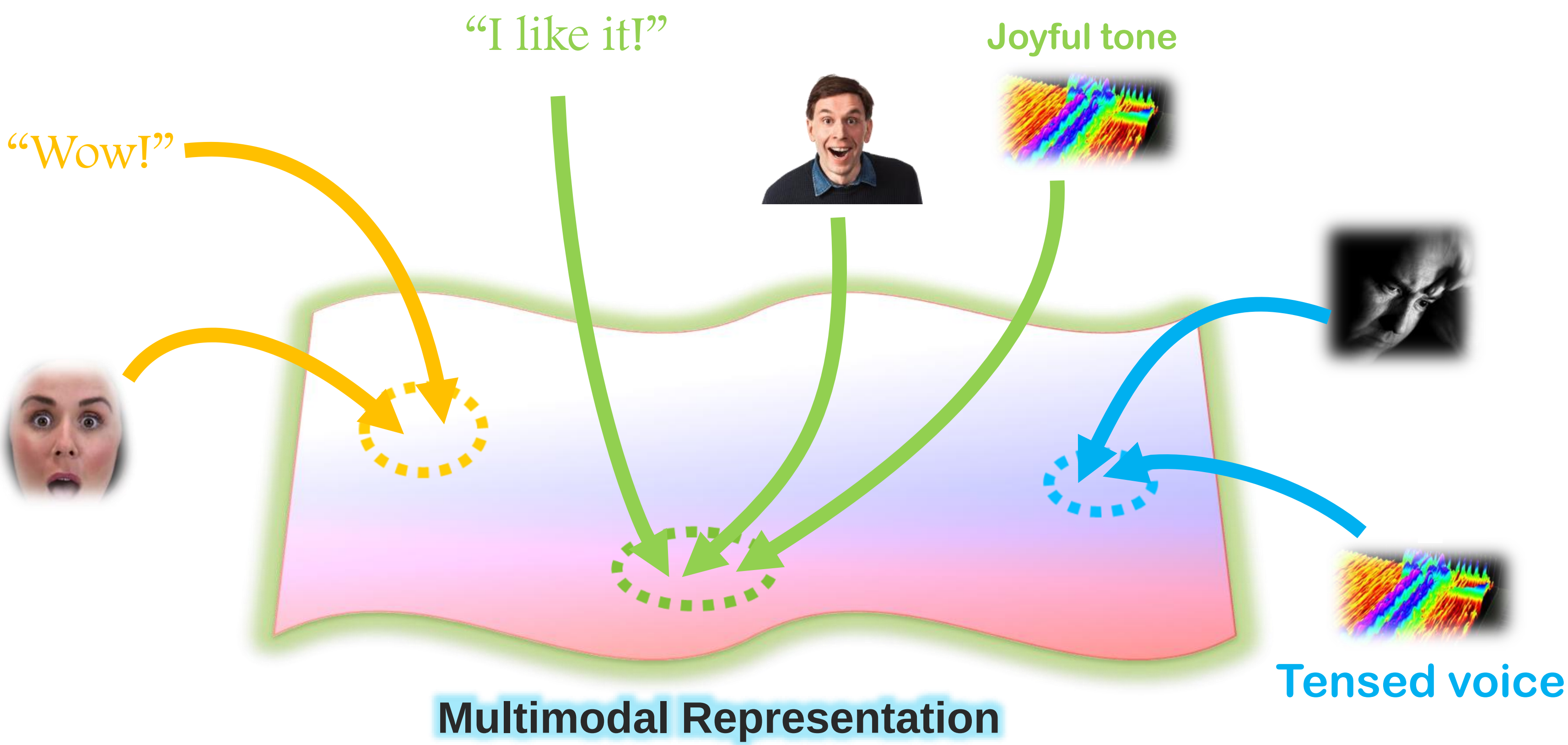
https://arxiv.org/abs/1705.09406

☑ **5 core challenges**
☑ **37 taxonomic classes**
☑ **253 referenced citations**

First Two Core Challenges



Core Challenge 1: Representation



Core Challenge 1: Early Examples

Audio-visual speech recognition

[Ngiam et al., ICML 2011]

- Bimodal Deep Belief Network

Image captioning

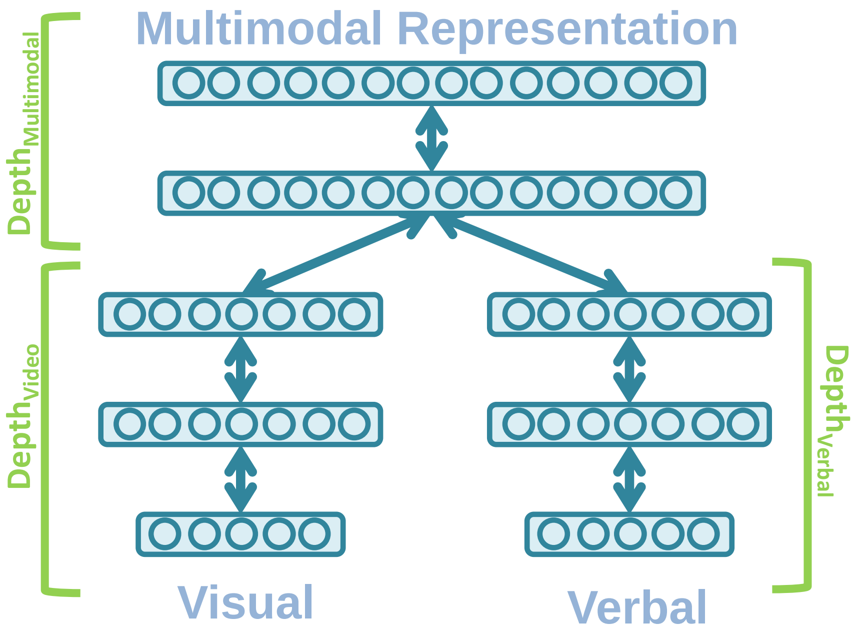
[Srivastava and Salahutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

Audio-visual emotion recognition

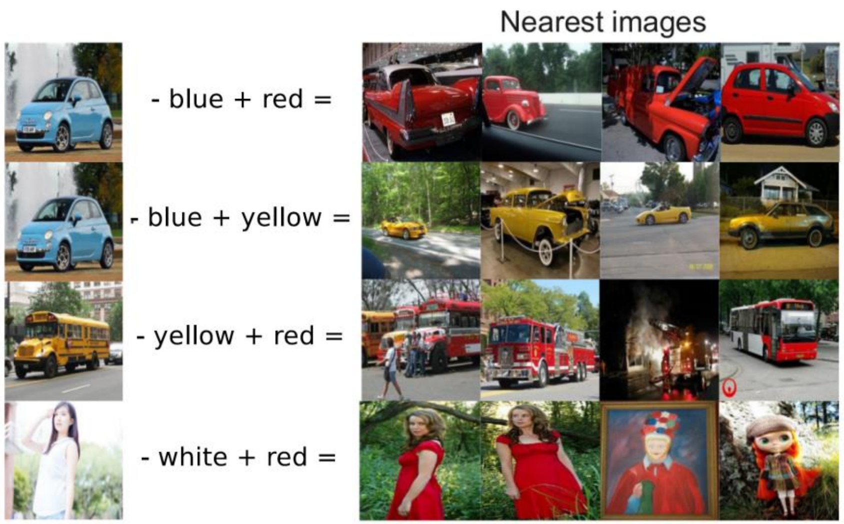
[Kim et al., ICASSP 2013]

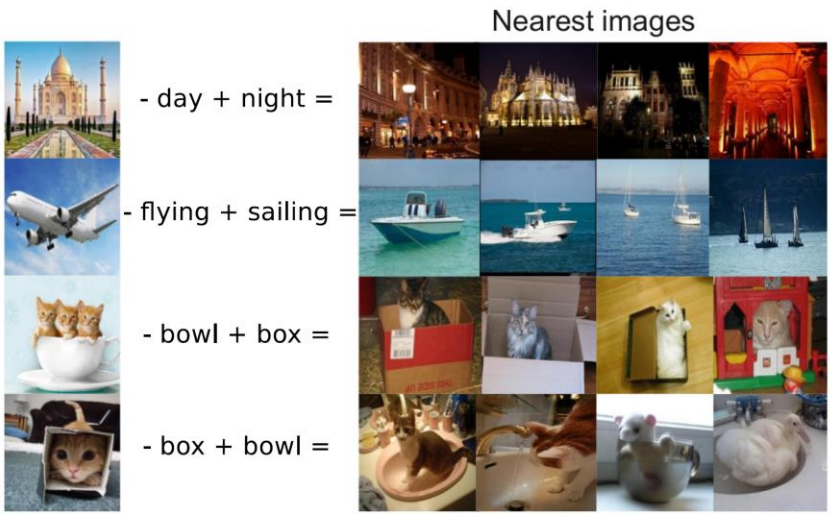
- Deep Boltzmann Machine



Core Challenge 1: Early Examples

Multimodal Vector Space Arithmetic



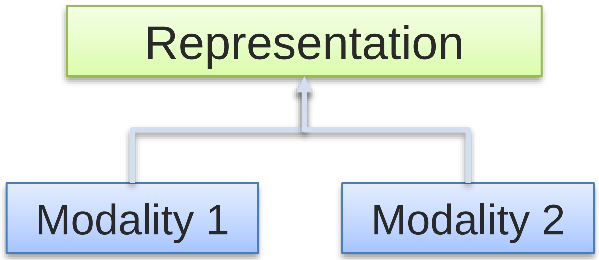


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

Core Challenge 1: Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

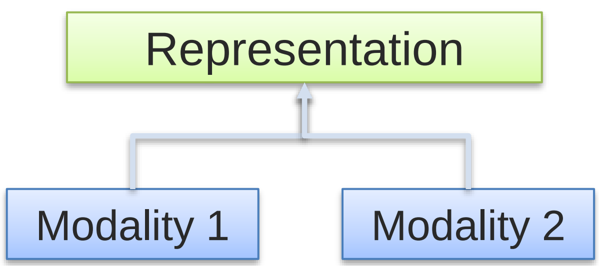
Ⓐ Joint representations:



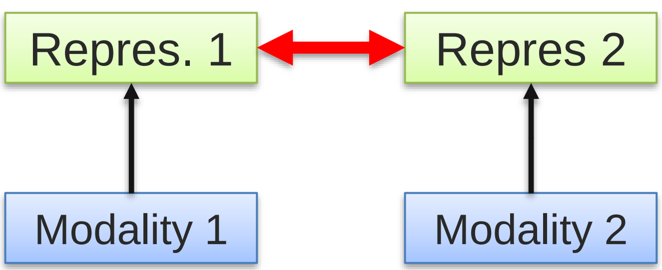
Core Challenge 1: Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

A Joint representations:

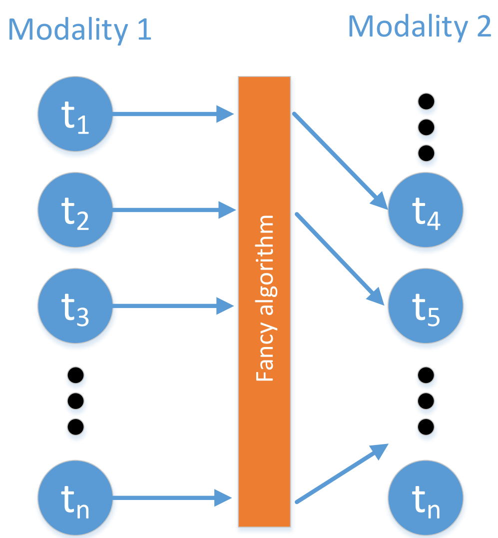


B Coordinated representations:



Core Challenge 2: Alignment

Definition: Identify the direct relations between (sub)elements from two or more different modalities.



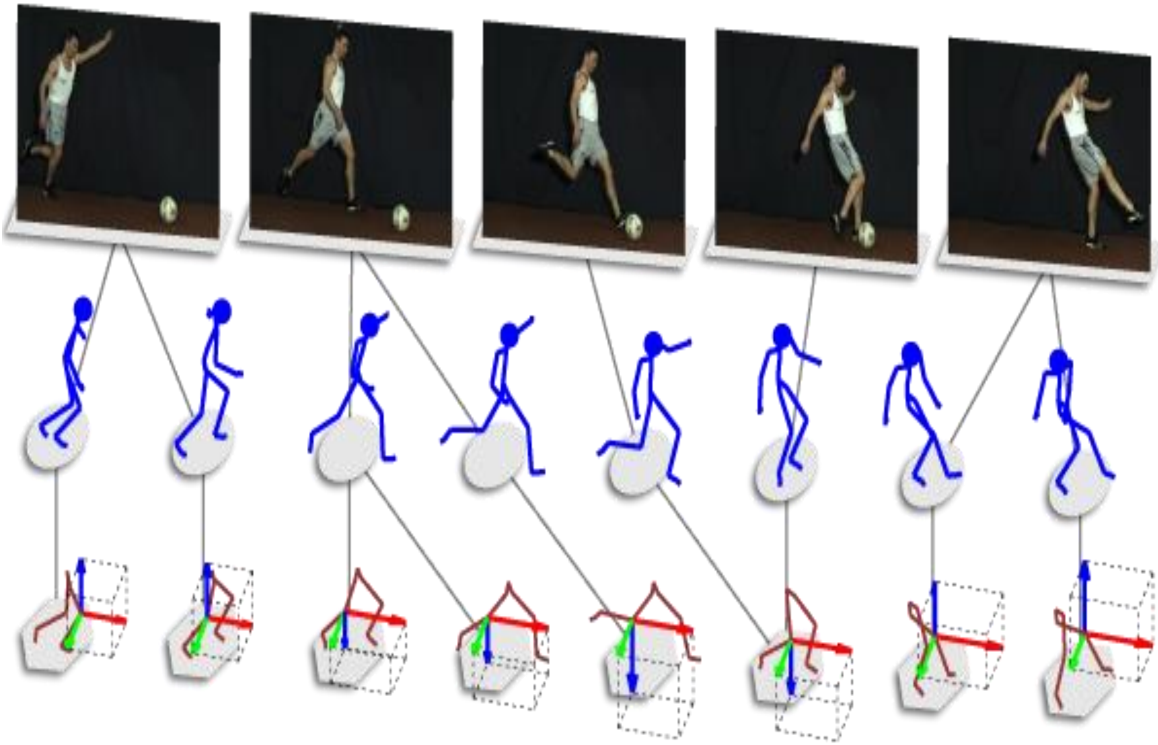
A Explicit Alignment

The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

Uses internally latent alignment of modalities in order to better solve a different problem

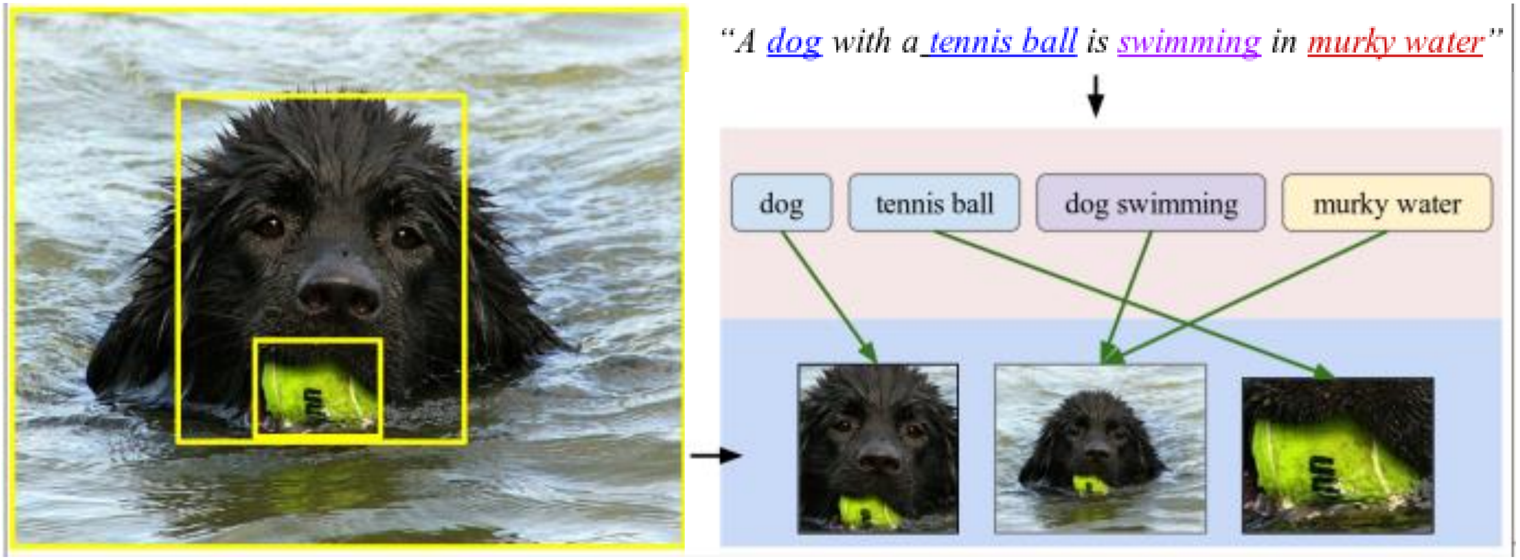
Core Challenge 2: Explicit Alignment



Applications:
- Re-aligning asynchronous data
- Finding similar data across modalities (we can estimate the aligned cost)
- Event reconstruction from multiple sources

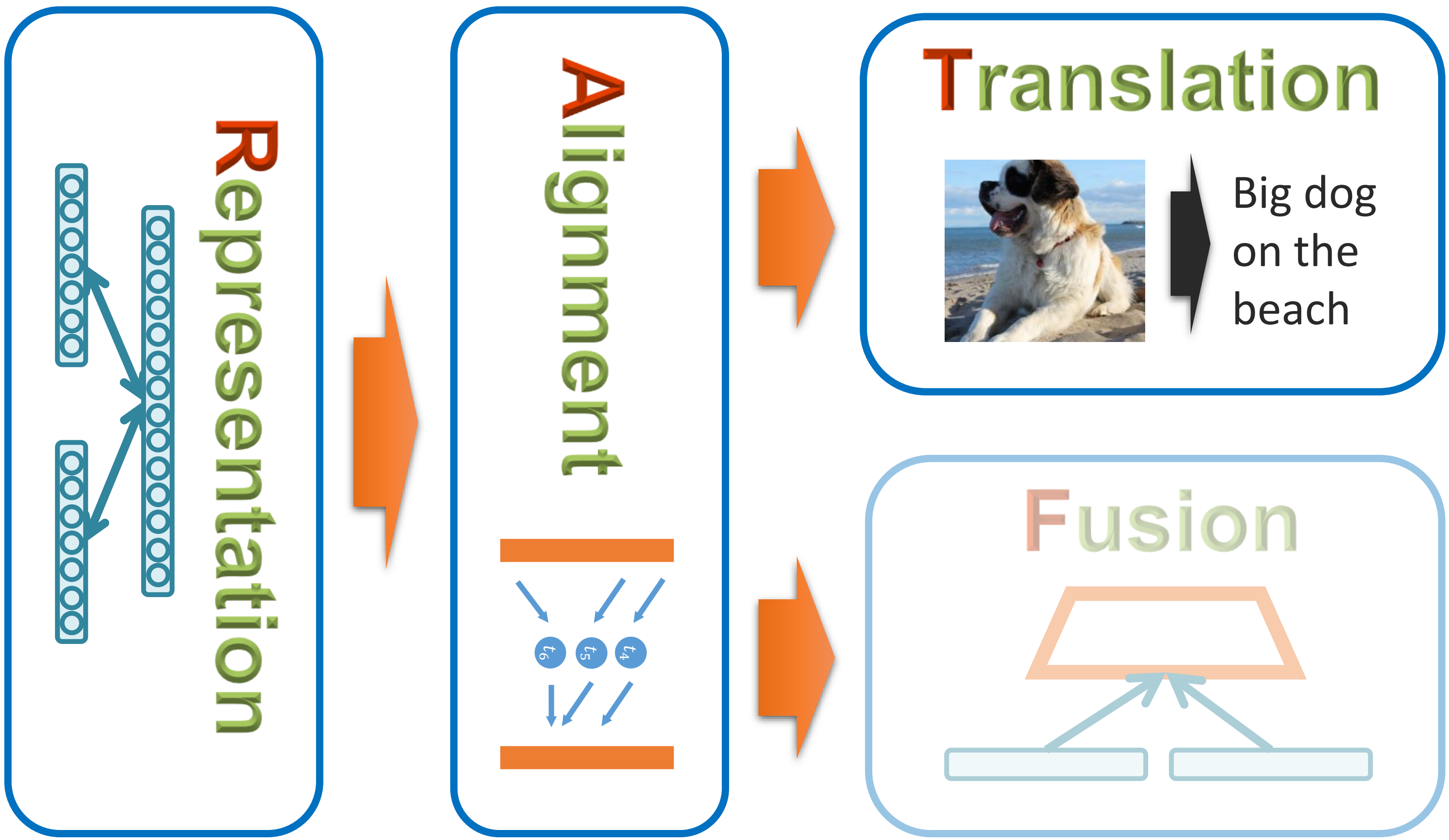
Core Challenge 2: Explicit Alignment

Core Challenge 2: Implicit Alignment



Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, https://arxiv.org/pdf/1406.5679.pdf

Two More Core Challenges



Core Challenge 3 – Translation

Visual gestures
(both speaker and listener gestures)

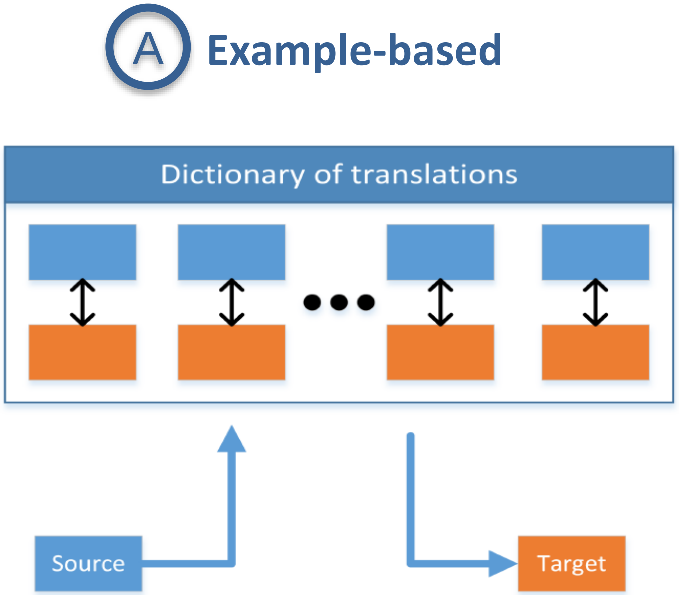
←

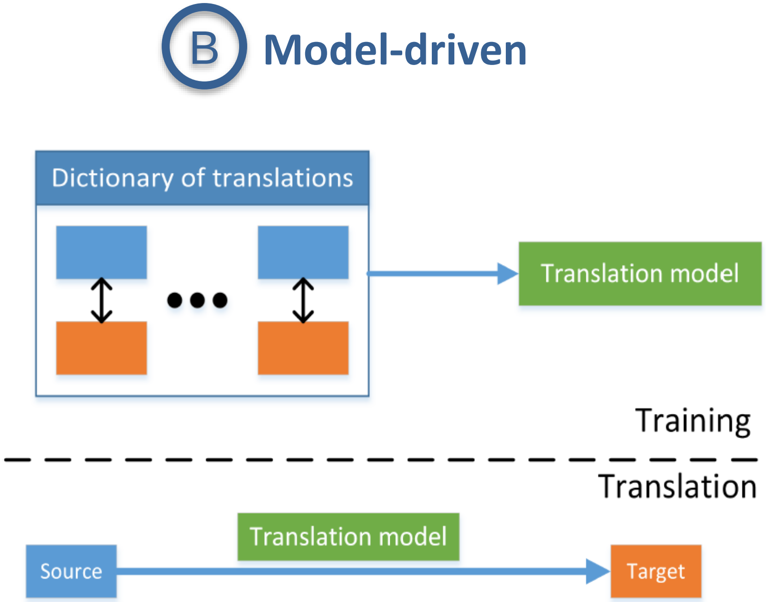
Transcriptions
+
Audio streams

Marsella et al., Virtual character performance from speech, SIGGRAPH/Eurographics Symposium on Computer Animation, 2013

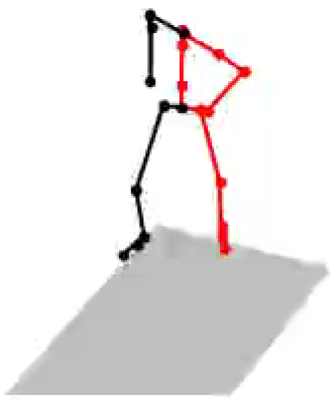
Core Challenge 3: Translation

Definition: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.

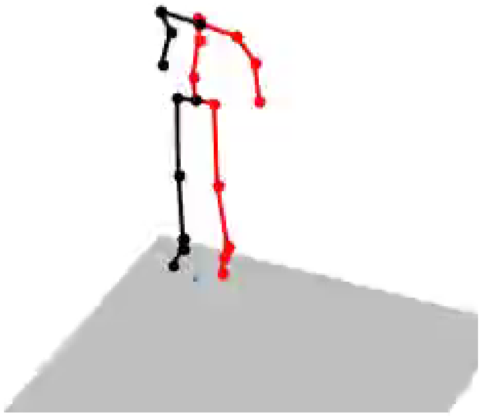




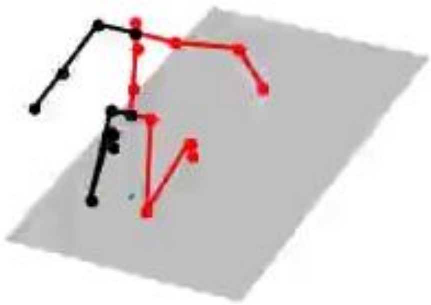
Core Challenge 3: Translation - Example



a person jogs a few steps



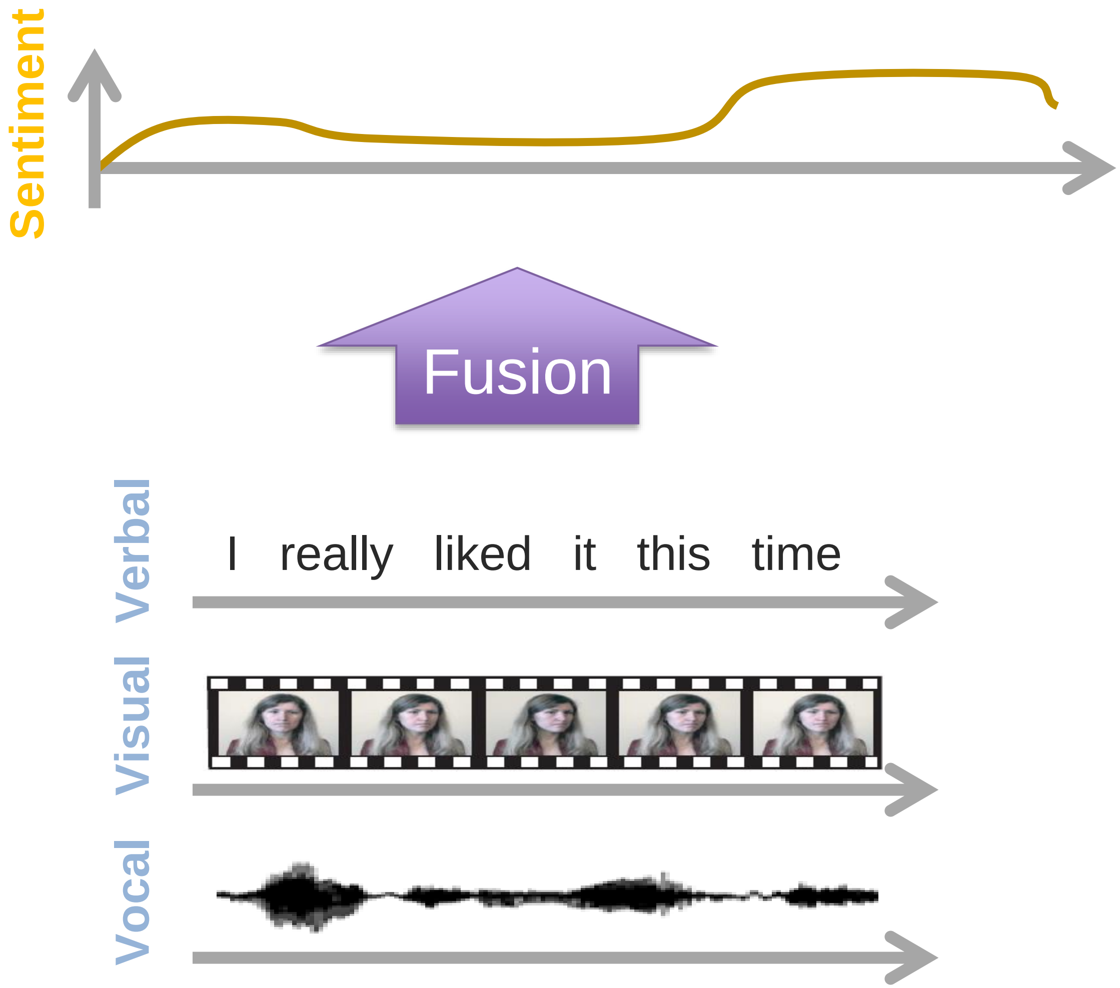
A person steps forward then turns around and steps forwards again.



A kneeling person raises their arms to the sides and stand up.

Ahuja, C., & Morency, L. P. (2019). Language2Pose: Natural Language Grounded Pose Forecasting. *Proceedings of 3DV Conference*

Core Challenge 4: Fusion

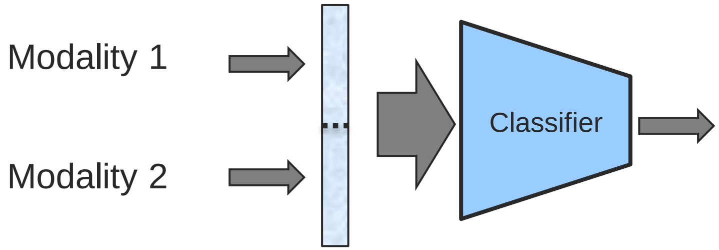


Core Challenge 4: Fusion

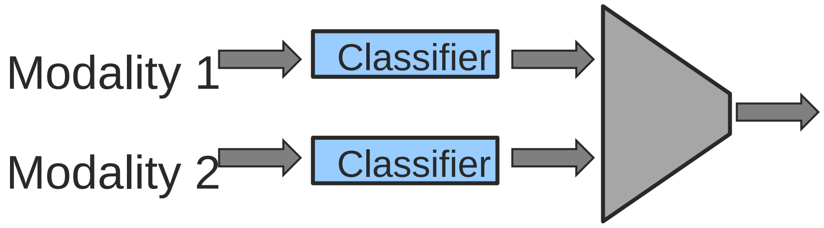
Definition: To join information from two or more modalities to perform a prediction task.

Ⓐ **Model-Agnostic Approaches**

1) Early Fusion



2) Late Fusion

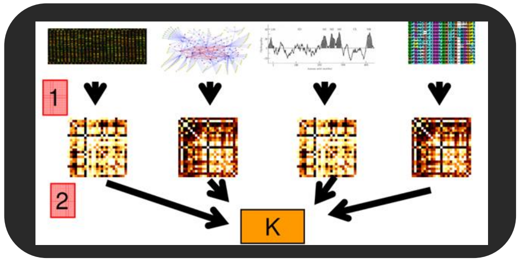


Core Challenge 4: Fusion

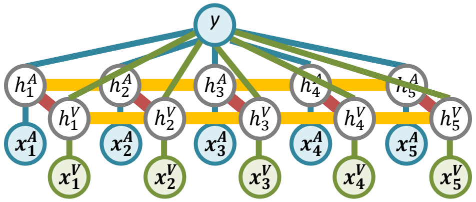
Definition: To join information from two or more modalities to perform a prediction task.

Ⓑ **Model-Based (Intermediate) Approaches**

1) **Deep neural networks**
2) **Kernel-based methods**
3) **Graphical models**

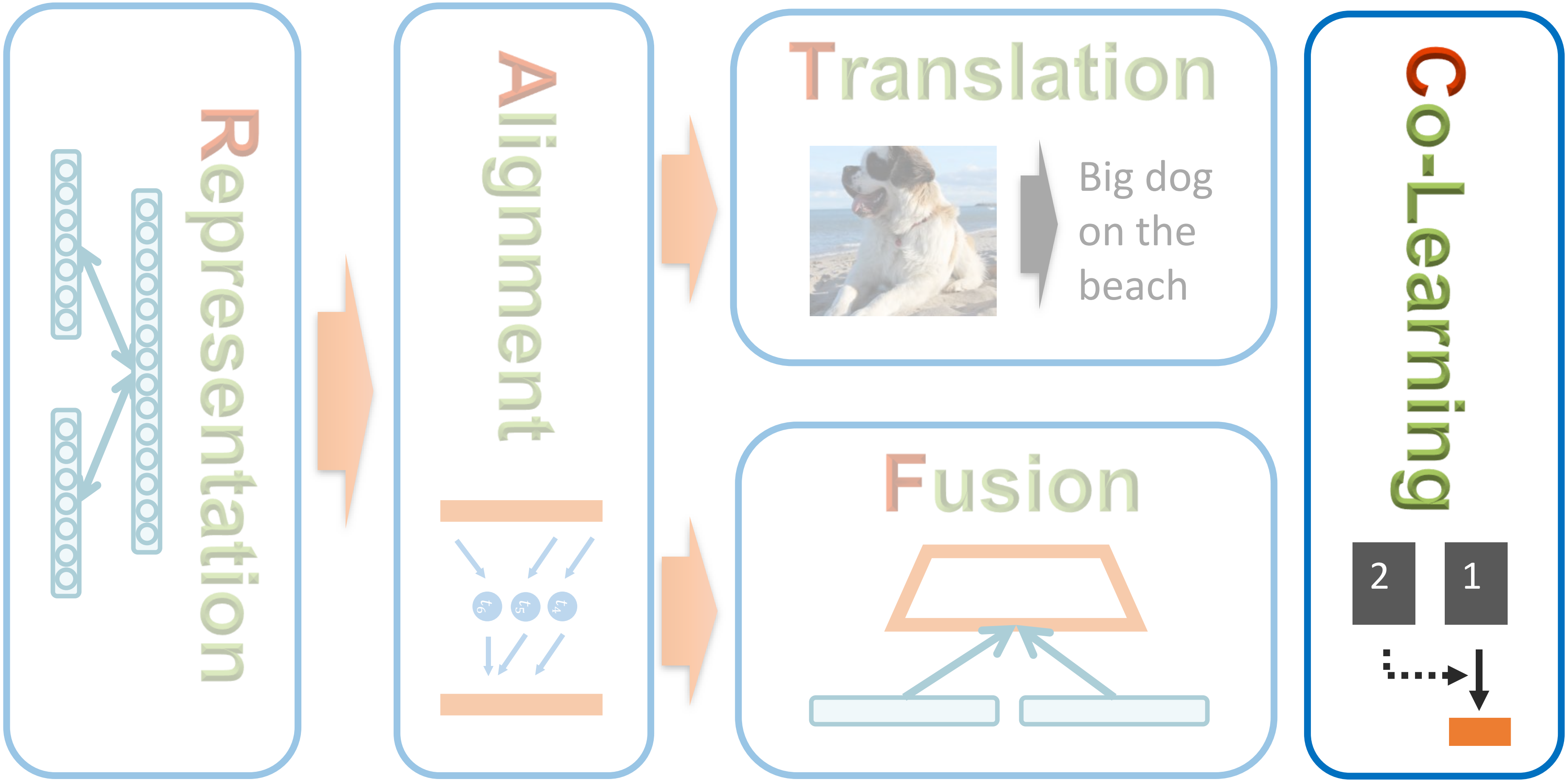


Multiple kernel learning



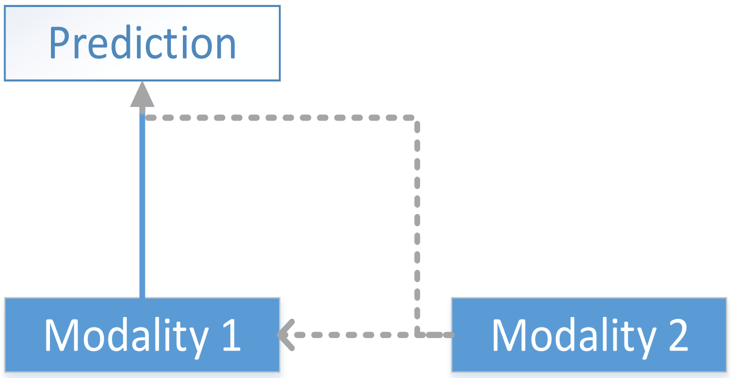
Multi-View Hidden CRF

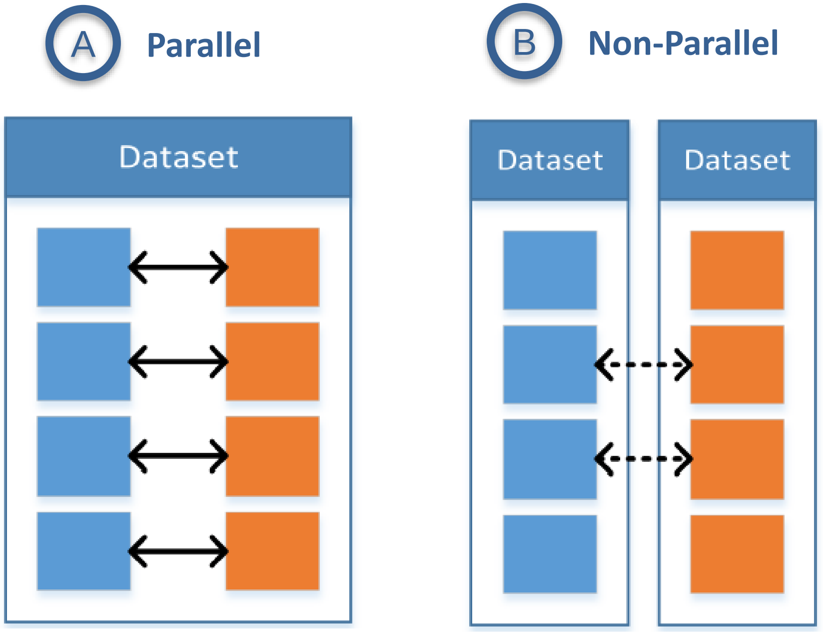
One Last Core Challenge



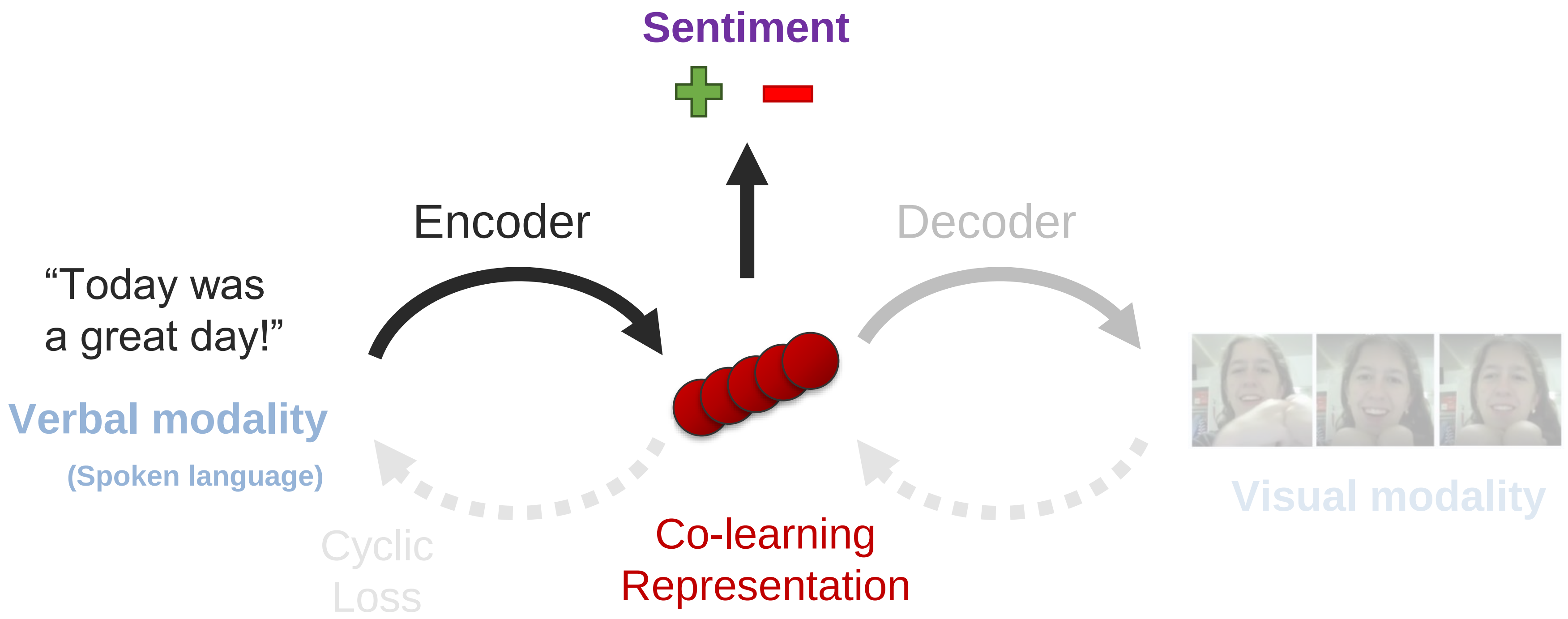
Core Challenge 5: Co-Learning

Definition: Transfer knowledge between modalities, including their representations and predictive models.



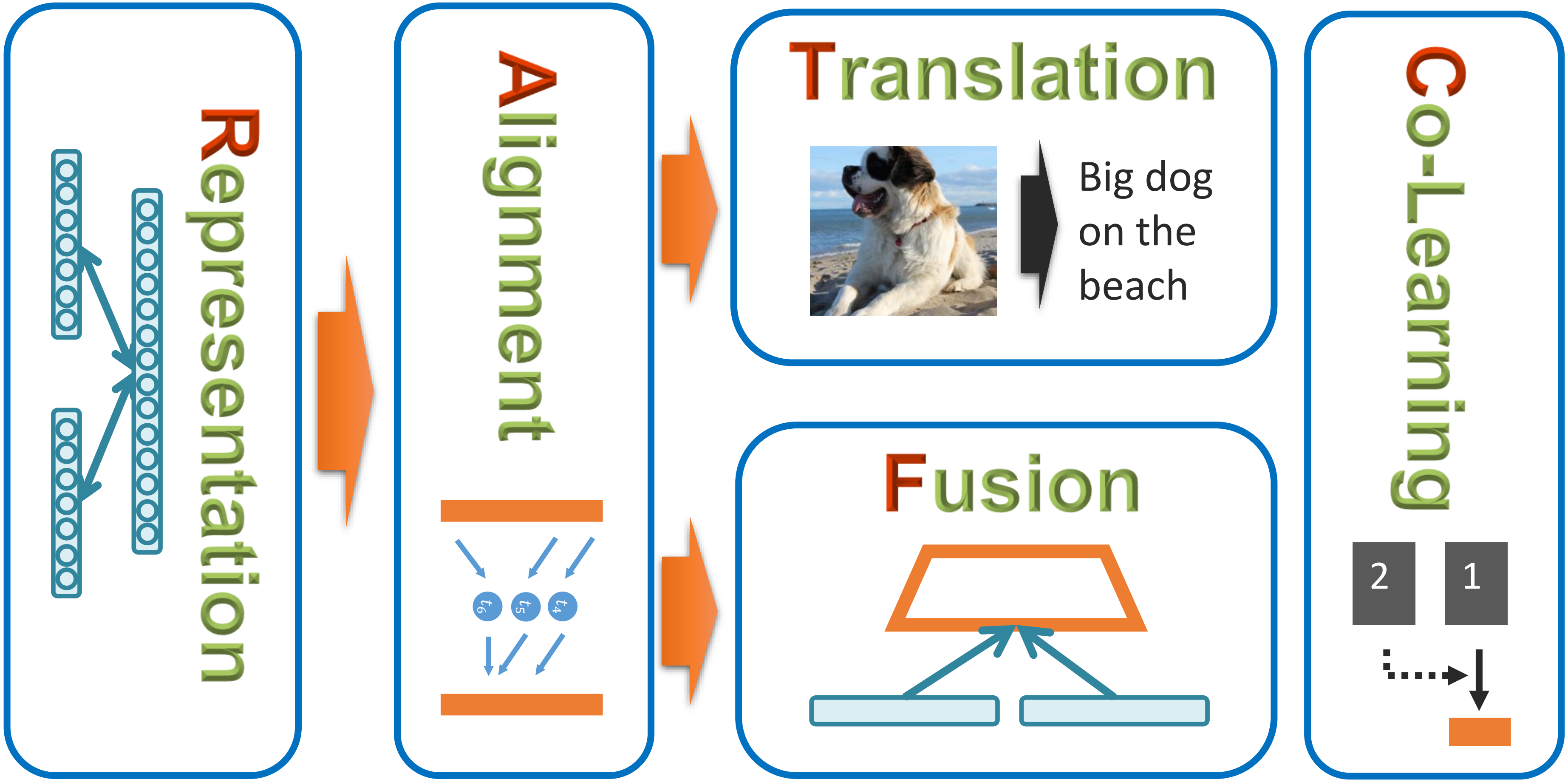


Core Challenge 5: Co-Learning



Pham et al., Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities, https://arxiv.org/abs/1812.07809

Five Multimodal Core Challenges



Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

Taxonomy of Multimodal Research

[https://arxiv.org/abs/1705.09406]

Representation

- Joint
  - *Neural networks*
  - *Graphical models*
  - *Sequential*
- Coordinated
  - *Similarity*
  - *Structured*

Translation

- Example-based
  - *Retrieval*
  - *Combination*
- Model-based
  - *Grammar-based*
  - *Encoder-decoder*
  - *Online prediction*

Alignment

- Explicit
  - *Unsupervised*
  - *Supervised*
- Implicit
  - *Graphical models*
  - *Neural networks*

Fusion

- Model agnostic
  - *Early fusion*
  - *Late fusion*
  - *Hybrid fusion*
- Model-based
  - *Kernel-based*
  - *Graphical models*
  - *Neural networks*

Co-learning

- Parallel data
  - *Co-training*
  - *Transfer learning*
- Non-parallel data
  - *Zero-shot learning*
  - *Concept grounding*
  - *Transfer learning*
- *Hybrid data*
  - *Bridging*

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

Real world tasks tackled by MMML

- Affect recognition
  - Emotion
  - Persuasion
  - Personality traits
- Media description
  - Image captioning
  - Video captioning
  - Visual Question Answering
- Event recognition
  - Action recognition
  - Segmentation
- Multimedia information retrieval
  - Content based/Cross-media

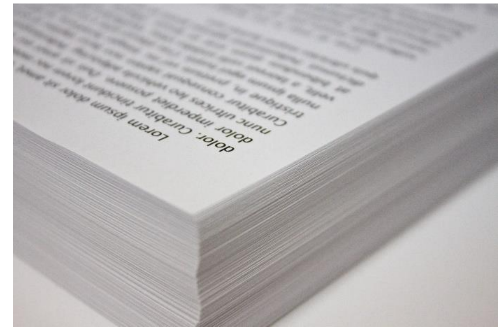
"man in black shirt is playing guitar." "construction worker in orange safety vest is working on road." "two young girls are playing with lego toy." "boy is doing backflip on wakeboard."

(a) answer-phone (a) get-out-car (a) fight-person (b) push-up (b) cartwheel

Course Syllabus

Three Course Learning Paradigms

Course lecture participation
(15% of your grade)



Reading assignments
(15% of your grade)

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)\\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)\\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

Course project assignments
(70% of your grade)

Course Recommendations and Requirements

1. Ready to read about 10 papers this semester !
   - Research papers as part of the weekly reading assignments
   - Summarize each paper and participate in group discussions
2. Already taken a machine learning course
   - Strongly recommended for students to have taken an introduction machine learning course
   - 10-401, 10-601, 10-701, 11-663, 11-441, 11-641 or 11-741
3. Motivated to produce a high-quality course project
   - Projects are designed to enhance state-of-the-art algorithms
   - Three project assignments, to help scaffold the project tasks

Course Project Timeline

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)\\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)\\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

- Pre-proposal (Wednesday 9/16)
  - Define your dataset, research task and teammates
- First project assignment (due Friday Oct. 9)
  - Experiment with unimodal representations
  - Study prior work on your selected research topic
- Midterm project assignment (due Friday Nov. 12)
  - Implement and evaluate state-of-the-art model(s)
  - Discuss new multimodal model(s)
- Final project assignment (due Friday Dec. 11)
  - Implement and evaluate new multimodal model(s)
  - Discuss results and possible future directions

Course Project Guidelines

- Dataset should have at least two modalities:
  - Natural language and visual/images
- Teams of 3, 4 or 5 students
- The project should explore algorithmic novelty
- Possible venues for your final report:
  - NAACL 2021, ACL 2021, IJCAI 2021, ICML 2021
- We will discuss on Thursday about project ideas
- GPU resources available:
  - Amazon AWS and Google Cloud Platform

Process for Selecting your Course Project

- **Thursday 9/3:** Lecture describing available multimodal datasets and research topics
- **Tuesday 9/8:** Let us know your dataset preferences for the course project
- **Thursday 9/10:** During the later part of the lecture, we will have an interactive period to help with team formation. More details to come
- **Wednesday 9/16:** Pre-proposals are due. You should have selected your teammates, dataset and task

# Equal Contribution by All Teammates!

- Each team will be required to create a GitHub repository which will be accessible by TAs
- Each report should include a description of the task from each teammate
- Please let us know soon if you have concerns about the participation levels of your teammates

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 1 9/1 & 9/3	**Course introduction** • Research and technical challenges • Course syllabus and requirements	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 9/8 & 9/10	**Basic concepts: neural networks** • Language, visual and acoustic • Loss functions and neural networks	**Basic concep** • Gradients a • Practical de
Week 3 9/15 & 9/17	**Visual unimodal representations** • Convolutional kernels and CNNs • Residual network and skip connection	**Language un** • Gated netw • Backpropag
Week 4 9/22 & 9/24	**Multimodal representation learning** • Multimodal auto-encoders • Multimodal joint representations	**Coordinated representations** • Deep canonical correlation analysis • Non-negative matrix factorization
Week 5 9/29 & 10/1	**Multimodal alignment** • Explicit - dynamic time warping • Implicit - attention models	**Structured representations** • Module networks • Tree-based and stack models
Week 6 10/6 & 10/8	***First project assignment*** *(live working sessions instead o*	

Project preferences due on Tuesday 9/8

Pre-proposals due on Wednesday 9/16

First assignment due on Friday 10/9

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 7 10/13 & 10/15	**Alignment and representation** • Multi-head attention • Multimodal transformers	**Probabilistic graphical models** • Dynamic Bayesian networks • Coupled and factor HMMs
Week 8 10/20 & 10/22	**Discriminative graphical models** • Conditional random fields • Continuous and fully-connected CRFs	**Neural Generative Models** • Variational auto-encoder • Generative adversarial networks
Week 9 10/27 & 10/29	**Reinforcement learning** • Markov decision process • Q learning and policy gradients	**Multimodal RL** • Deep Q learning • Multimodal applications
Week 10 11/3 & 11/5	**Fusion and co-learning** • Multi-kernel learning and fusion • Few shot learning and co-learning	**New research directions** • Recent approaches in multimodal ML
Week 11 11/10 & 11/12	***Mid-term project assignment*** *(live working sessions*	

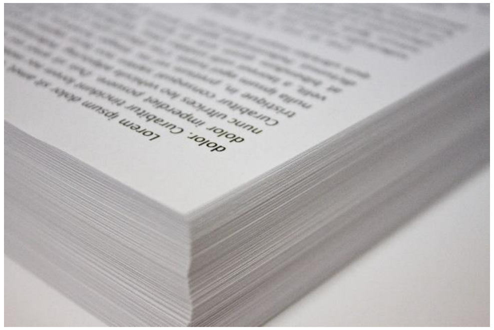
Midterm due on 11/12.

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 12 11/17 & 11/19	**Embodied Language Grounding** • Connecting Language to Action • Guest lecture: Yonatan Bisk	**Multi-lingual representations** • Tentative topic • Guest lecture: To be confirmed
Week 13 11/24 & 11/26	***Thanksgiving week*** *(no lectures)*	
Week 14 12/1 & 12/3	**Bias and fairness** • Tentative topic • Guest lecture: To be confirmed	**Learning to connect text and images** • Discourse approaches, text & images • Guest lecture: Malihe Alikhani
Week 15 12/8 & 12/10	***Final project assignment*** *(live working sessions inste*	

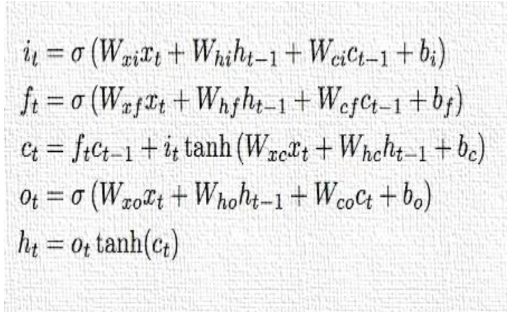
Final due on 12/11.

Course Grades



- Lecture participation 16%
- Reading assignments 16%

- Project preferences/pre-proposal 3%
- First project assignment
  - Report and presentation 15%
- Mid-term project assignment
  - Report and presentation 20%
- Final project assignment
  - Report and presentation 30%



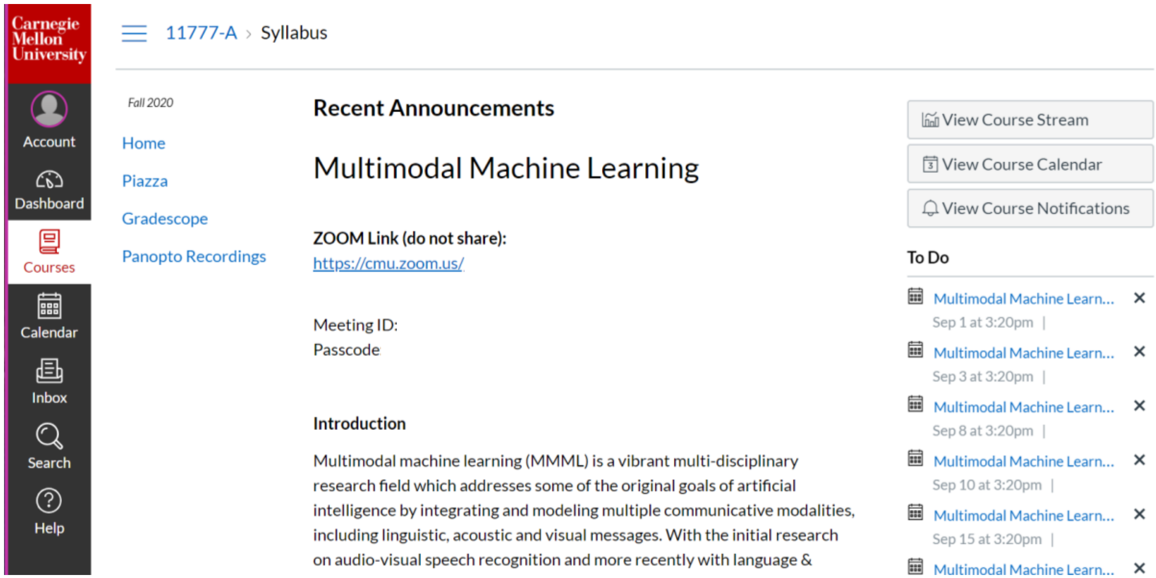
Lecture Participation – Highlight Forms

- Students should summarize lecture highlights
  - Each lecture is split in 3 segments (~30mins each)
  - One highlight statement for each segment
    - This is the main takeaway from this segment
  - Optionally, students can include related question
- Highlights submitted 42 hours after the lecture
  - Lecture can be watched live or asynchronously
- Questions will be summarized by TAs
  - Answers posted on Piazza

Reading Assignments

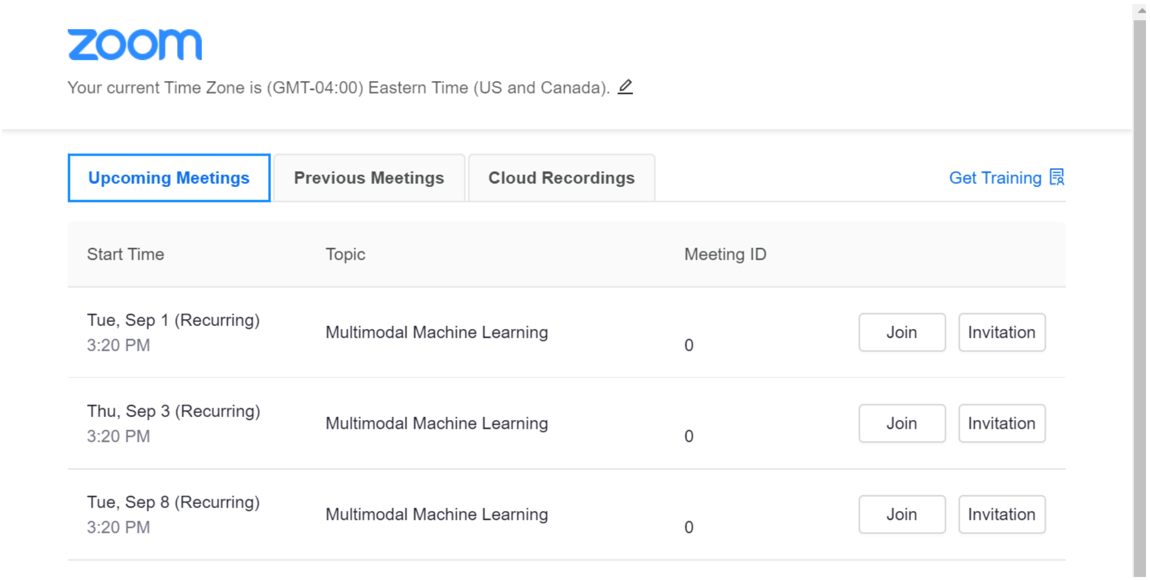
- 3 papers for each reading assignment
  - **Each student will read only one paper!**
  - Then you will create a short summary to help others
- Discussions with your study group
  - 9-10 students in each study group
  - Discuss together the 3 papers. Ask questions!
    - But you should also try to answer the questions
- Graded based on summary and discussion
  - 1 point for the summary and 1 point for the discussion

Canvas https://canvas.cmu.edu/courses/18106



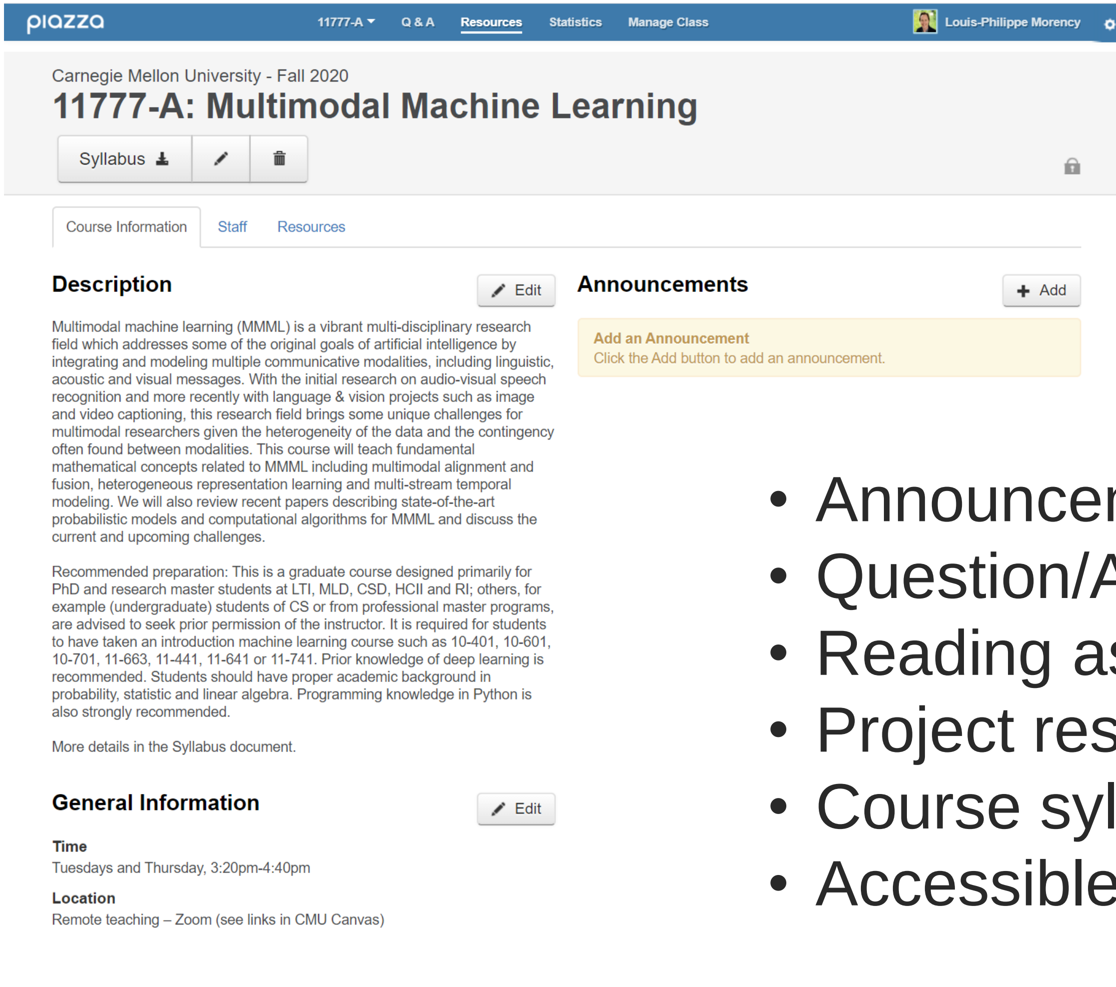
- Main launching pad for everything related to the course
    - Zoom, Piazza, Gradescope
    - Recorded lectures on Panopto
- Course syllabus

Zoom & Panopto



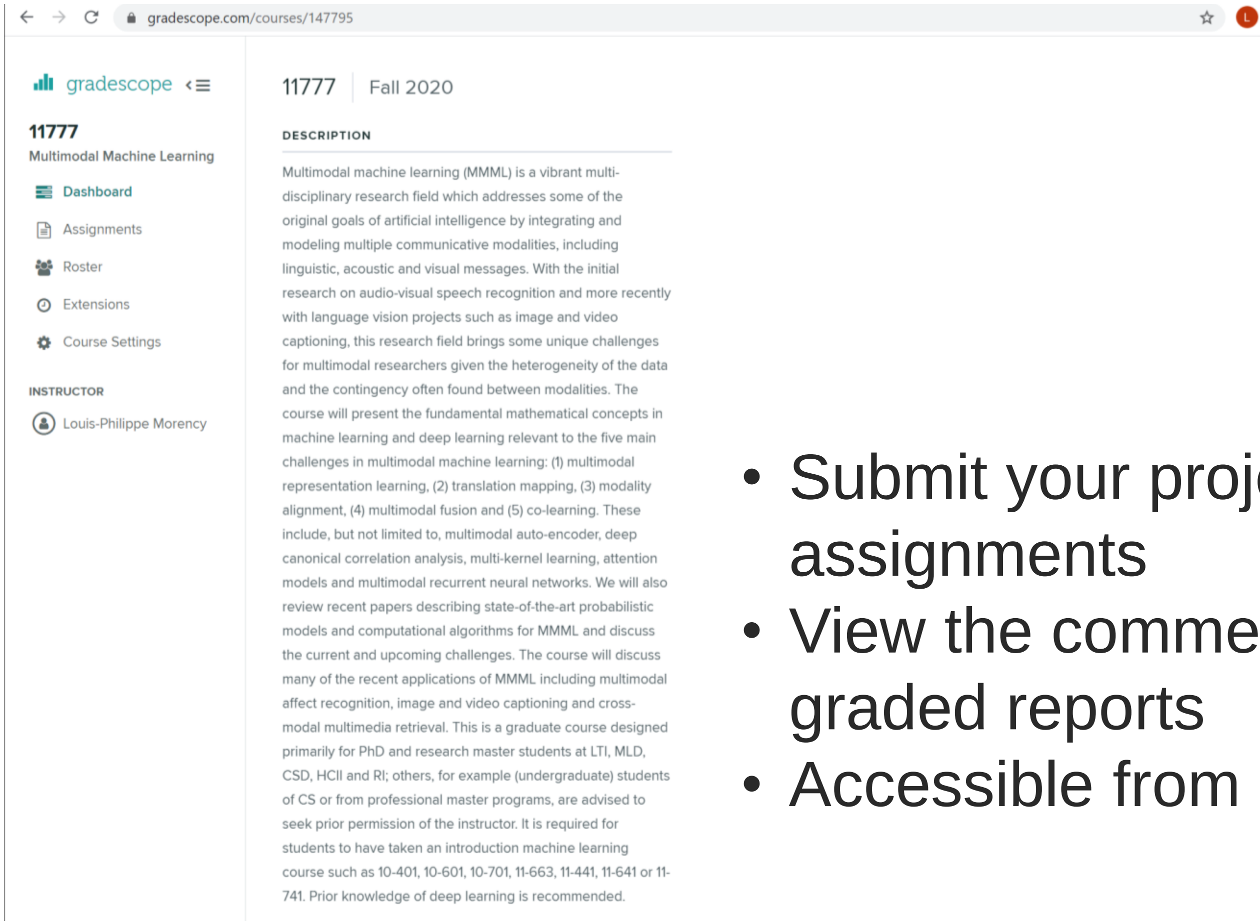
- Live lectures (with Zoom)
- Recorded lectures (with Panopto)
- Links accessible from Canvas

Piazza https://piazza.com/cmu/fall2020/11777/home



- Announcements
- Question/Answers
- Reading assignments
- Project resources
- Course syllabus
- Accessible from Canvas

Gradescope


gradescope.com/courses/147795

gradescope

11777

Multimodal Machine Learning

Dashboard

Assignments

Roster

Extensions

Course Settings

INSTRUCTOR

Louis-Philippe Morency

11777 | Fall 2020

DESCRIPTION

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including linguistic, acoustic and visual messages. With the initial research on audio-visual speech recognition and more recently with language vision projects such as image and video captioning, this research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities. The course will present the fundamental mathematical concepts in machine learning and deep learning relevant to the five main challenges in multimodal machine learning: (1) multimodal representation learning, (2) translation mapping, (3) modality alignment, (4) multimodal fusion and (5) co-learning. These include, but not limited to, multimodal auto-encoder, deep canonical correlation analysis, multi-kernel learning, attention models and multimodal recurrent neural networks. We will also review recent papers describing state-of-the-art probabilistic models and computational algorithms for MMML and discuss the current and upcoming challenges. The course will discuss many of the recent applications of MMML including multimodal affect recognition, image and video captioning and cross-modal multimedia retrieval. This is a graduate course designed primarily for PhD and research master students at LTI, MLD, CSD, HCII and RI; others, for example (undergraduate) students of CS or from professional master programs, are advised to seek prior permission of the instructor. It is required for students to have taken an introduction machine learning course such as 10-401, 10-601, 10-701, 11-663, 11-441, 11-641 or 11-741. Prior knowledge of deep learning is recommended.

- Submit your project assignments
- View the comments from your graded reports
- Accessible from Canvas

External Course Website

11-777 MMML   logistics schedule homework project reports

MultiModal Machine Learning

11-777 • Fall 2020 • Carnegie Mellon University

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including linguistic, acoustic, and visual messages. With the initial research on audio-visual speech recognition and more recently with language & vision projects such as image and video captioning, this research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities. This course will teach fundamental mathematical concepts related to MMML including multimodal alignment and fusion, heterogeneous representation learning and multistream temporal modeling. We will also review recent papers describing state-of-the-art probabilistic models and computational algorithms for MMML and discuss the current and upcoming challenges.

The course will present the fundamental mathematical concepts in machine learning and deep learning relevant to the five main challenges in multimodal machine learning: (1) multimodal representation learning, (2) translation & mapping, (3) modality alignment, (4) multimodal fusion and (5) co-learning. These include, but not limited to, multimodal auto-encoder, deep canonical correlation analysis, multi-kernel learning, attention models and multimodal recurrent neural networks. The course will also discuss many of the recent applications of MMML including multimodal affect recognition, image and video captioning and cross-modal multimedia retrieval.

- Public link of recorded lectures (with some delays)
- List of reading assignments
- List of final project videos (this is optional)

https://cmu-multicomp-lab.github.io/mmml-course/fall2020/

Spring 2021 Edition of the MMML Course !

Yonatan Bisk
ybisk@cs.cmu.edu
https://yonatanbisk.com/

More details about the Spring edition to come later!

Project Preferences – Due Tuesday 9/8

- Post your project preferences:
  - List of your ranked preferred projects
    - Use alphanumeric code of each dataset
    - Detailed dataset list in the "Lecture1.2-datasets“ slides
  - Previous unimodal/multimodal experience
  - Available CPU / GPU resources
- For topics or datasets not in the list:
  - Include a description with links (for other students)

https://piazza.com/cmu/fall2020/11777/home