Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 1.1: Introduction

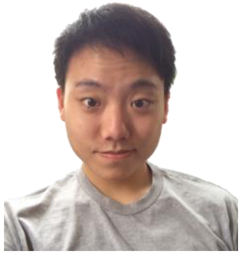
Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Your Instructor and TAs This Semester (11-777)

Louis-Philippe Morency
morency@cs.cmu.edu
Office: GHC-5411

Jonathan Francis
jmf1@andrew.cmu.edu



Paul Liang
pliang@andrew.cmu.edu

Sai Krishna Rallabandi
srallaba@andrew.cmu.edu

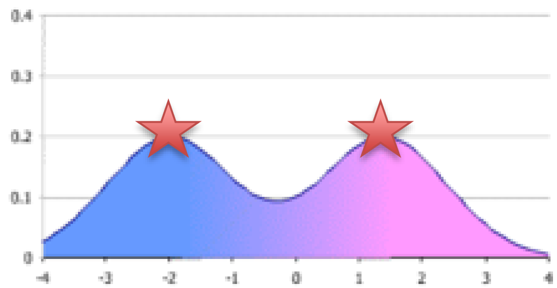
Ying Shen
yshen2@andrew.cmu.edu

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?

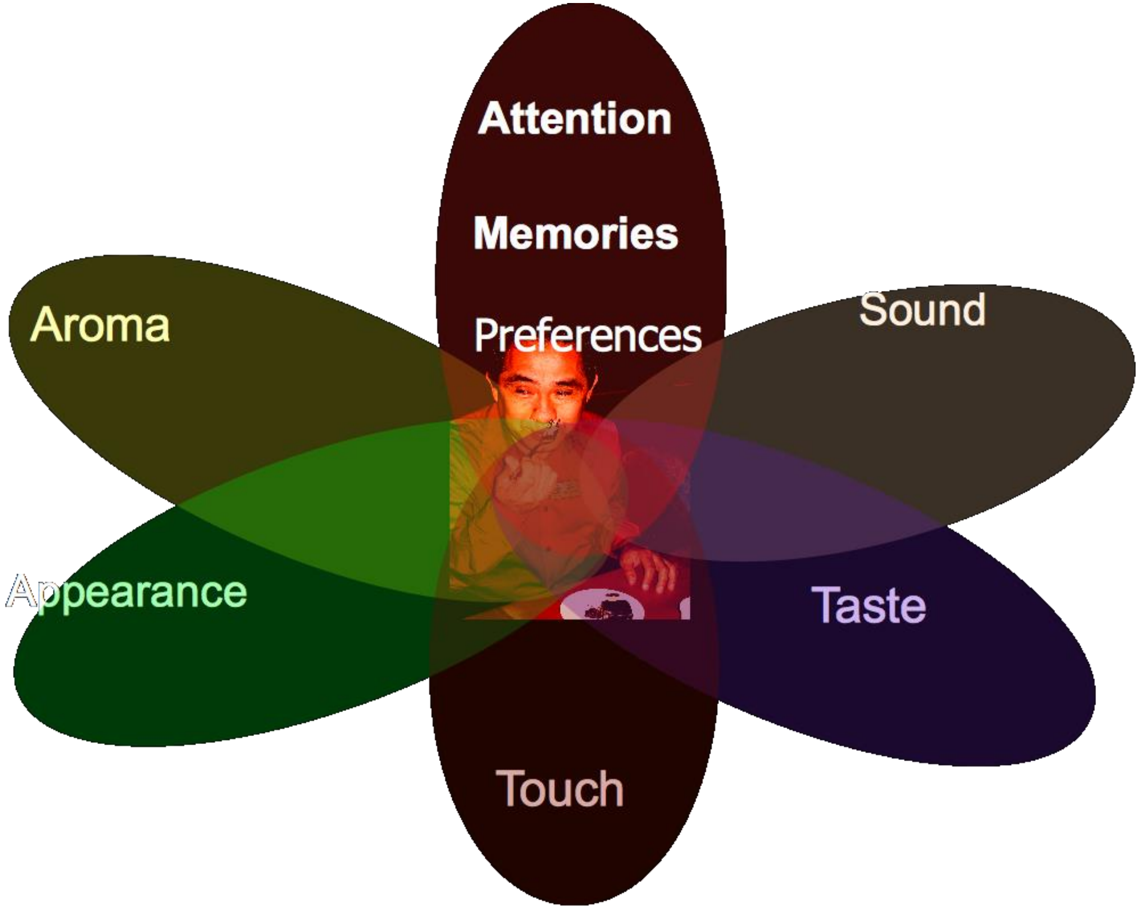
What is Multimodal?



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.

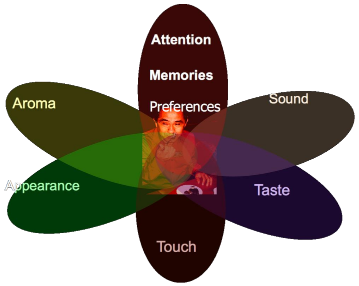
- *Modality* refers to a certain type of information and/or the representation format in which information is stored.
- *Sensory modality:* one of the primary forms of sensation, as vision or touch; channel of communication.

Medium (“middle”)

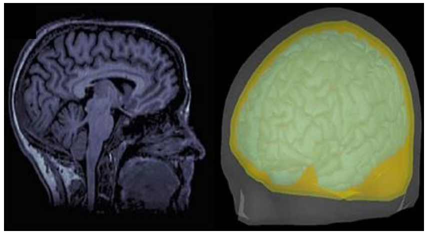
A means or instrumentality for storing or communicating information; system of communication/transmission.

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

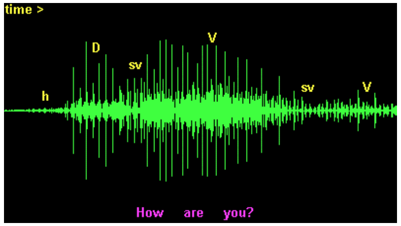
Multiple Communities and Modalities



Psychology



Medical

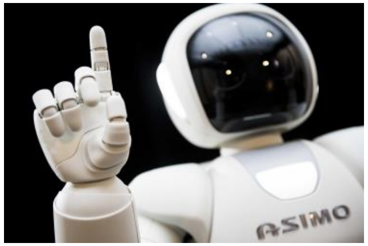


Speech

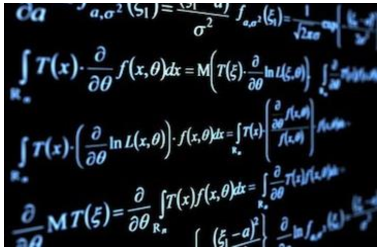
Vision

Language

Multimedia



Robotics



Learning

Examples of Modalities

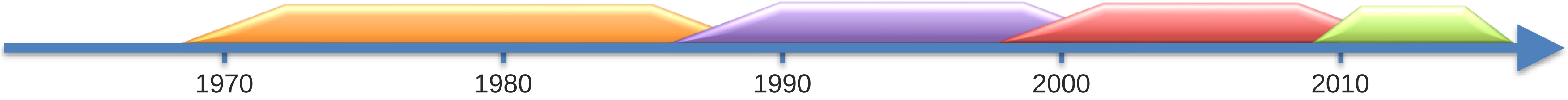
- Natural language (both spoken or written)
- Visual (from images or videos)
- Auditory (including voice, sounds and music)
- Haptics / touch
- Smell, taste and self-motion
- Physiological signals
  - Electrocardiogram (ECG), skin conductance
- Other modalities
  - Infrared images, depth images, fMRI

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

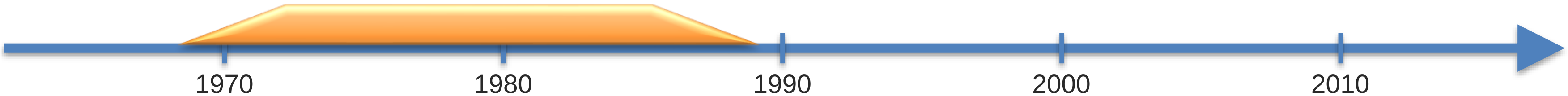


Language and Gestures

David McNeill
University of Chicago
Center for Gesture and Speech Research

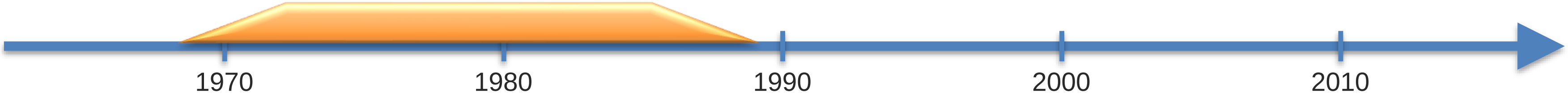
“For McNeill, gestures are in effect the speaker’s thought in action, and integral components of speech, not merely accompaniments or additions.”

- TRIVIA: Justine Cassell was a student of David McNeill



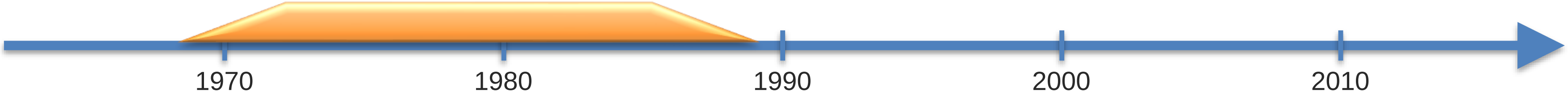
The McGurk Effect (1976)

[Hearing lips and seeing voices – Nature]



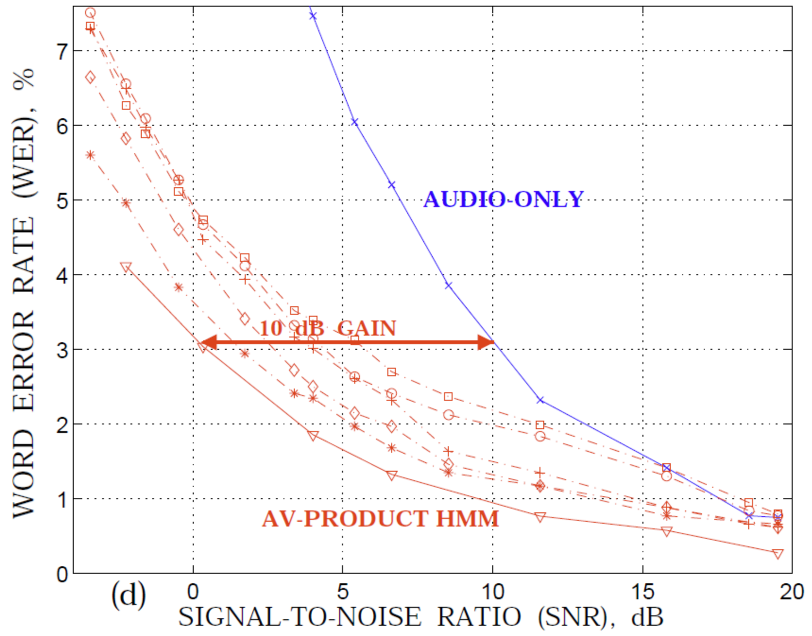
The McGurk Effect (1976)

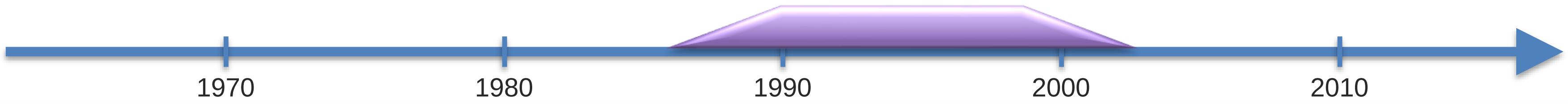
[Hearing lips and seeing voices – Nature]



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

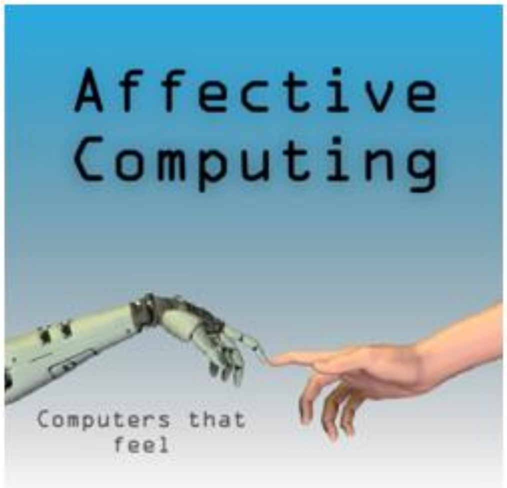
1) Audio-Visual Speech Recognition (AVSR)





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

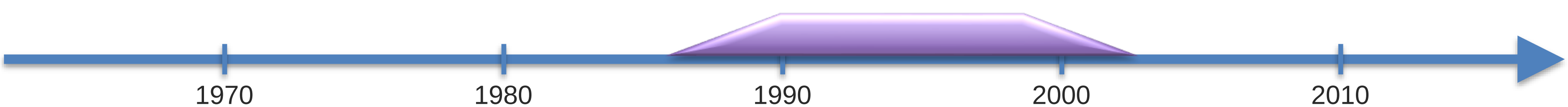
2) Multimodal/multisensory interfaces



Rosalind Picard

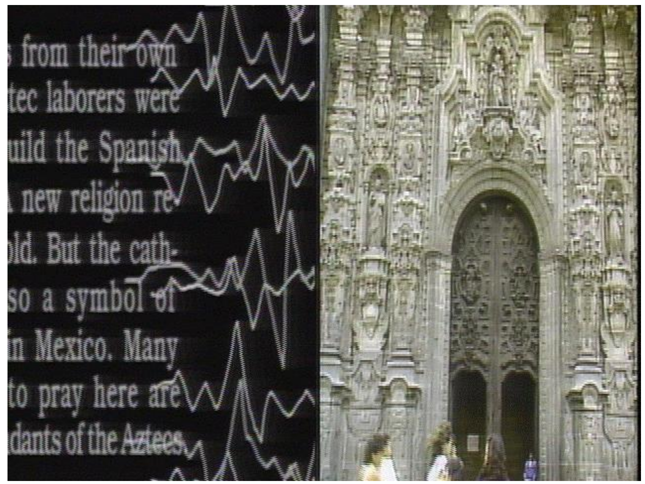
Affective Computing *is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena.*

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



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

3) Multimedia Computing

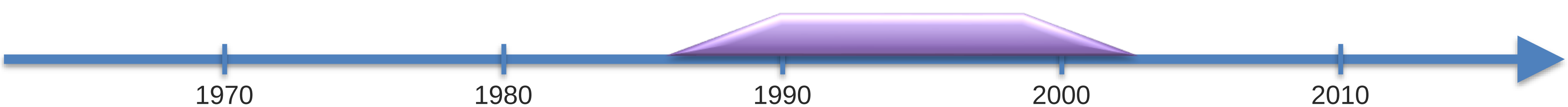


Carnegie Mellon University

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

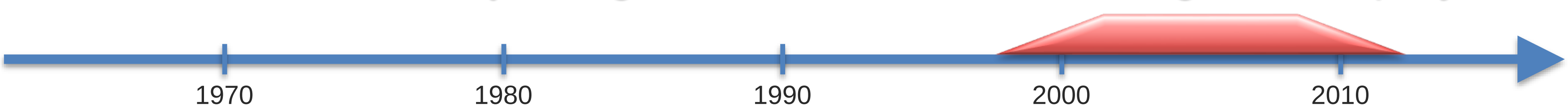
AMI Project [2001-2006, IDIAP]

- 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

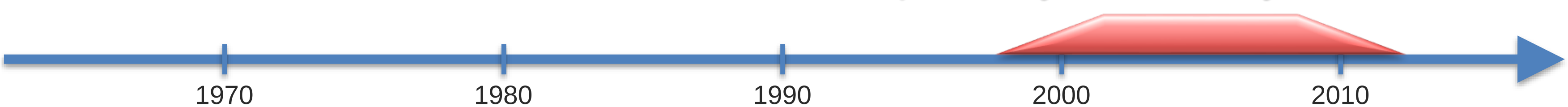
CALO Project [2003-2008, SRI]

- 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

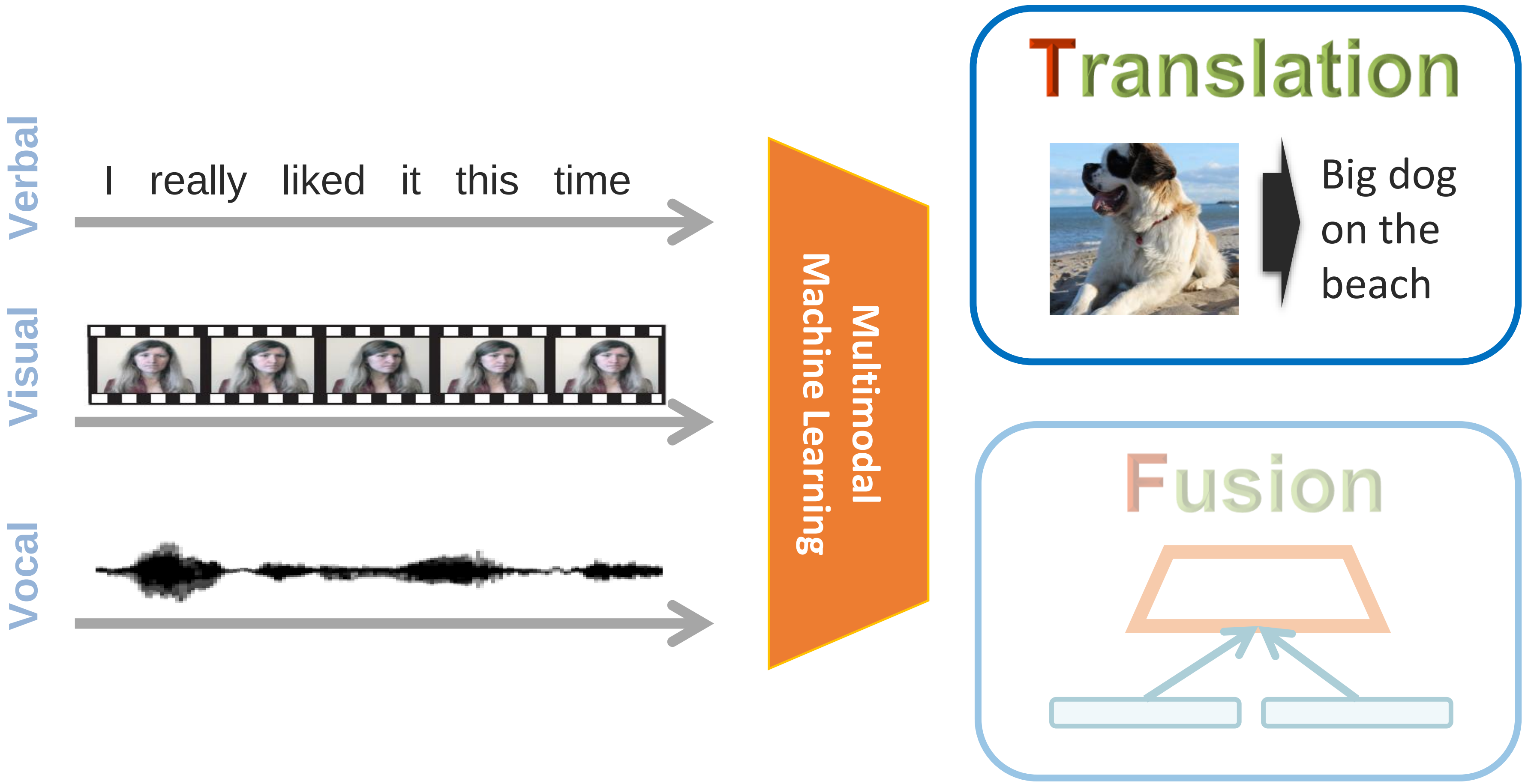
Multimodal Machine Learning: A Survey and Taxonomy

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 – Translation

Visual gestures
(both speaker and listener gestures)

←

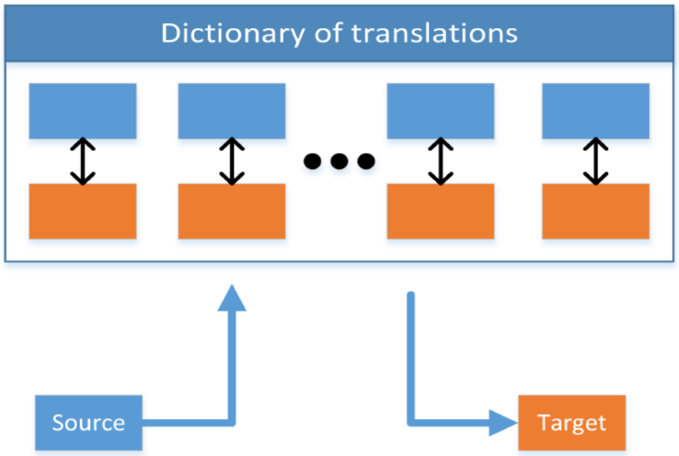
Transcriptions + Audio streams

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

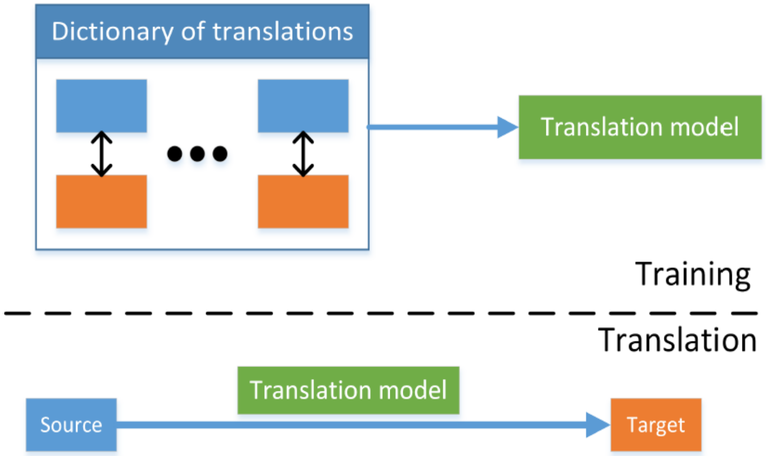
Core Challenge 1: Translation

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

A **Example-based**



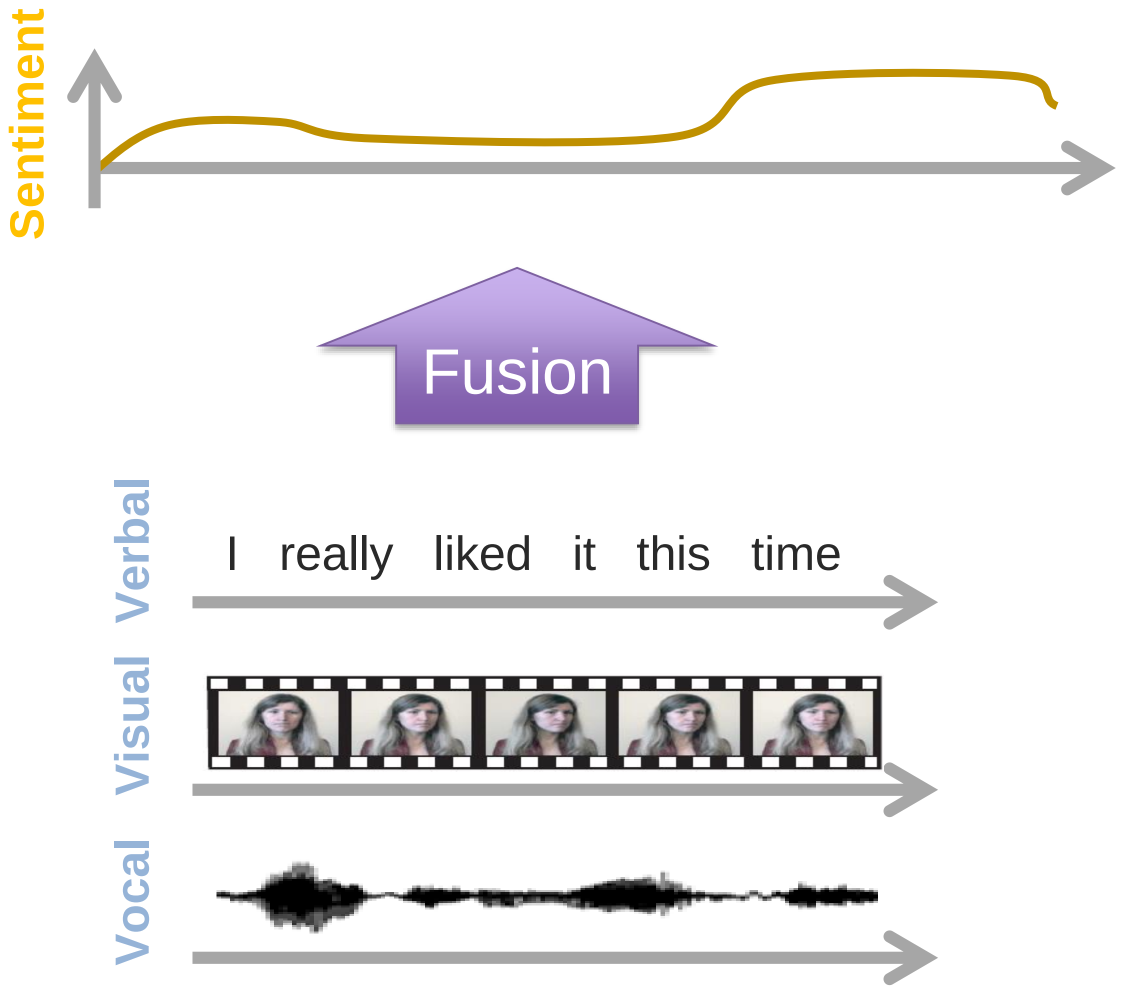
B **Model-driven**



Core Challenge 1: Translation - Example

Ahuja, C., & Morency, L. P. (2019). Language2Pose: Natural Language Grounded Pose Forecasting. *arXiv preprint arXiv:1907.01108*.

Core Challenge 2: Fusion

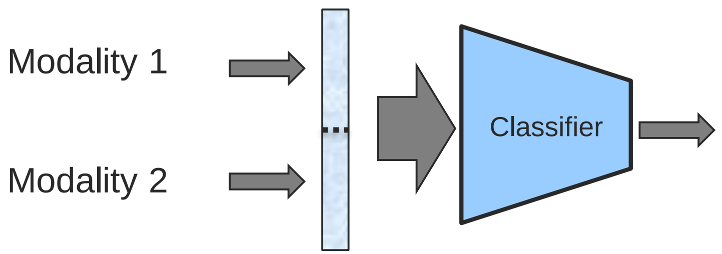


Core Challenge 2: Fusion

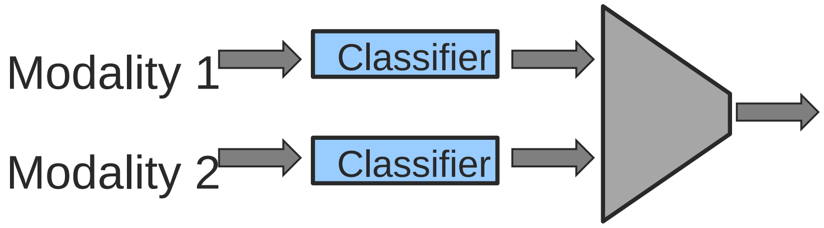
Definition: To join information from two or more modalities to perform a prediction task.

Ⓐ Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion

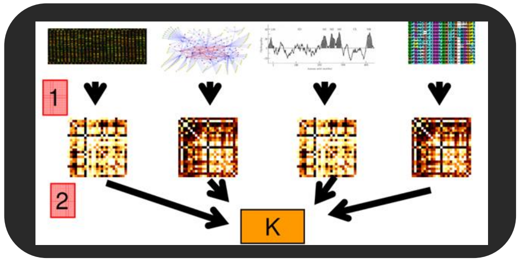


Core Challenge 2: Fusion

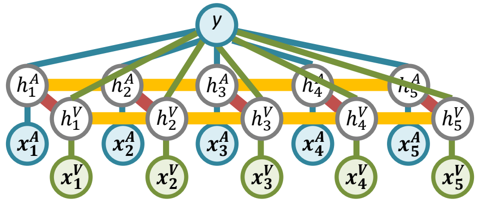
Definition: To join information from two or more modalities to perform a prediction task.

B Model-Based (Intermediate) Approaches

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

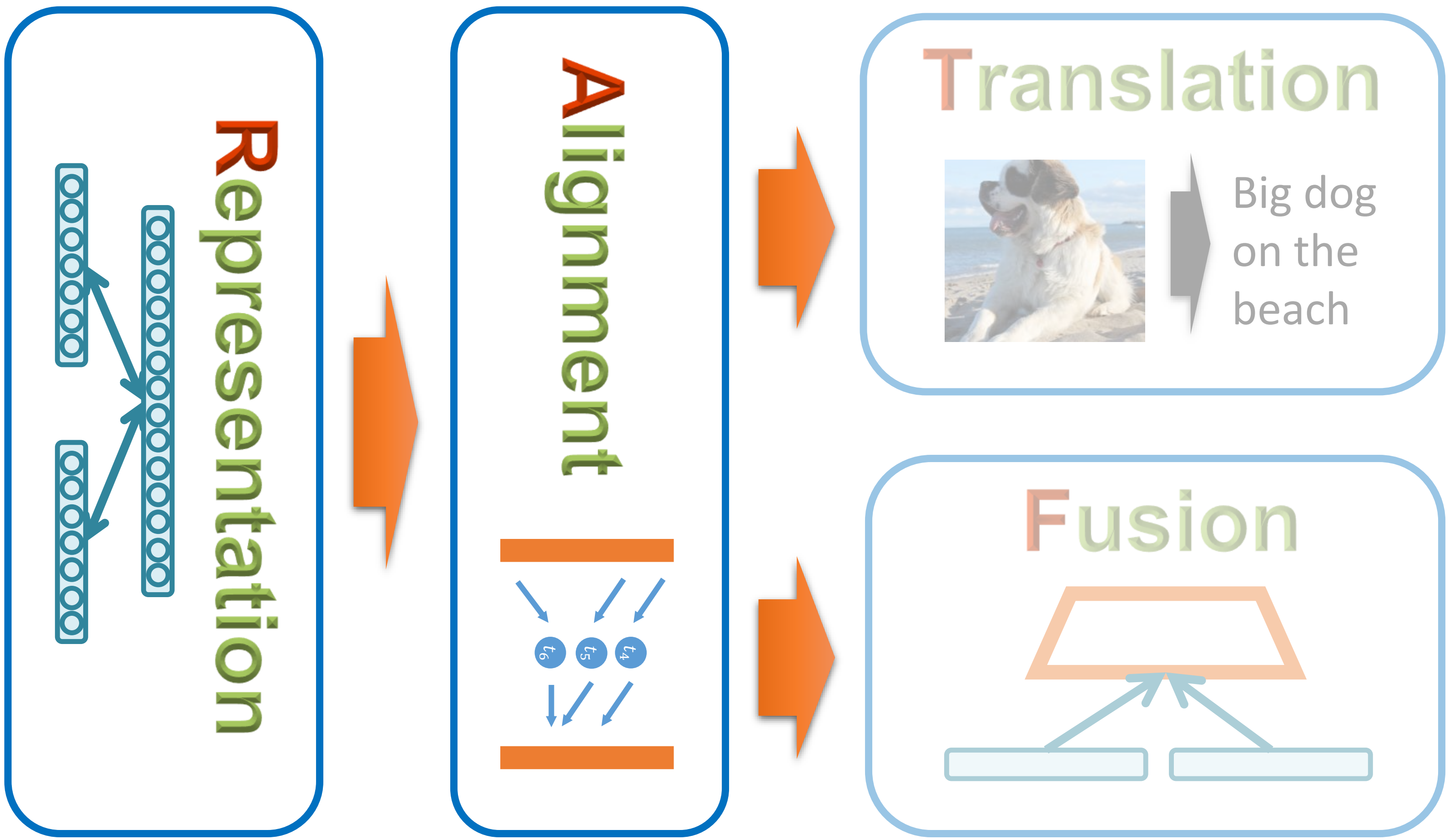


Multiple kernel learning

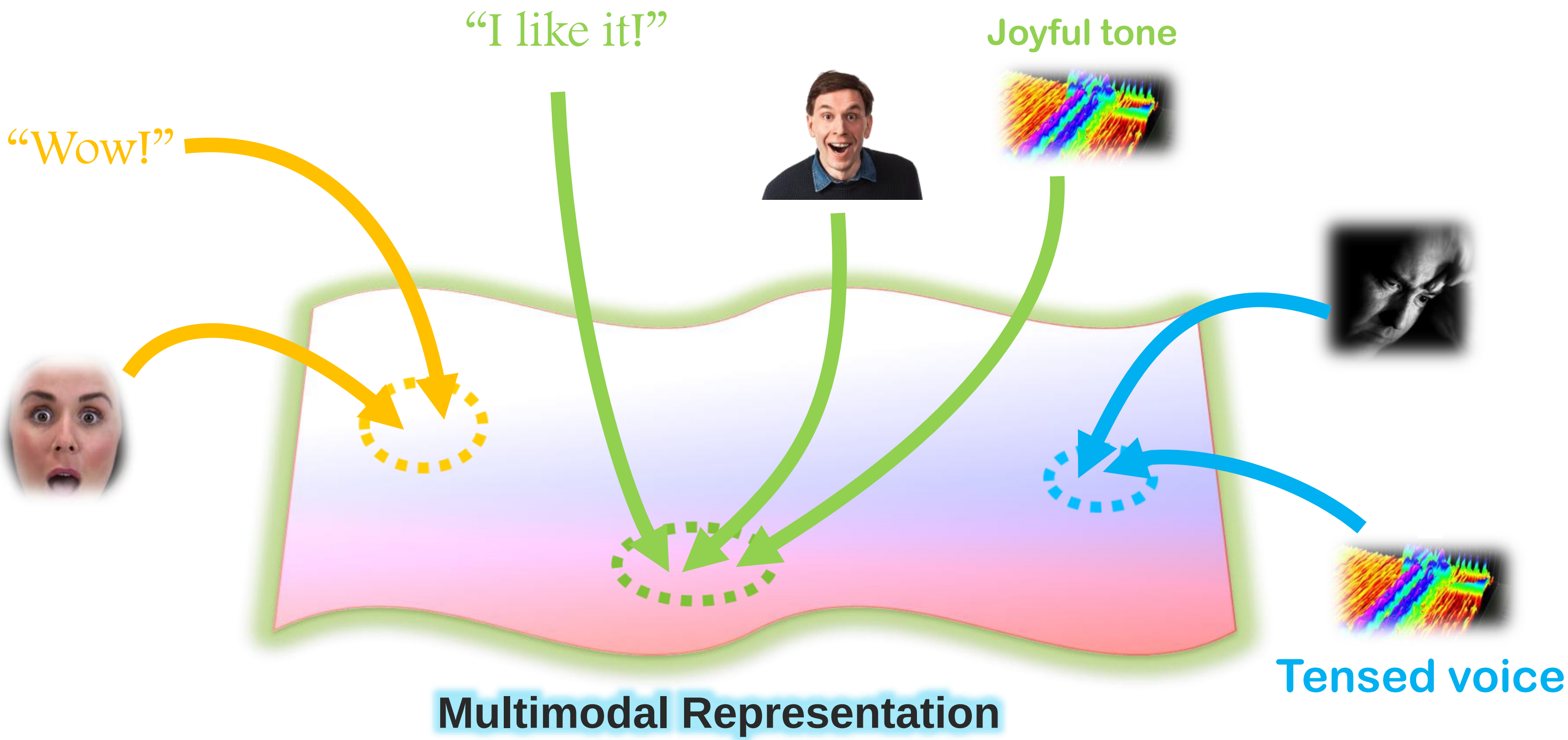


Multi-View Hidden CRF

Two More Core Challenges



Core Challenge 3: Representation



Core Challenge 3: Early Examples

Audio-visual speech recognition

[Ngiam et al., ICML 2011]

- Bimodal Deep Belief Network

Image captioning

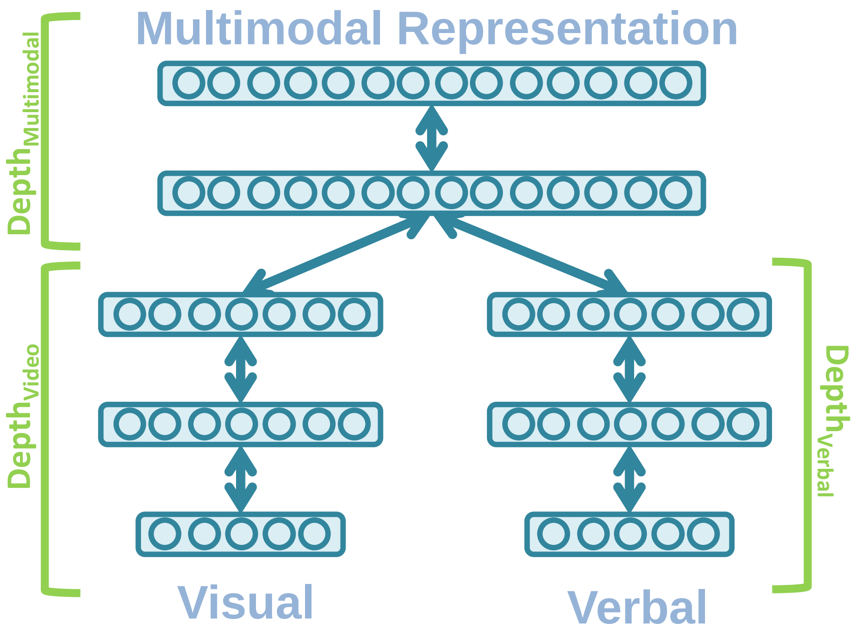
[Srivastava and Salahutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

Audio-visual emotion recognition

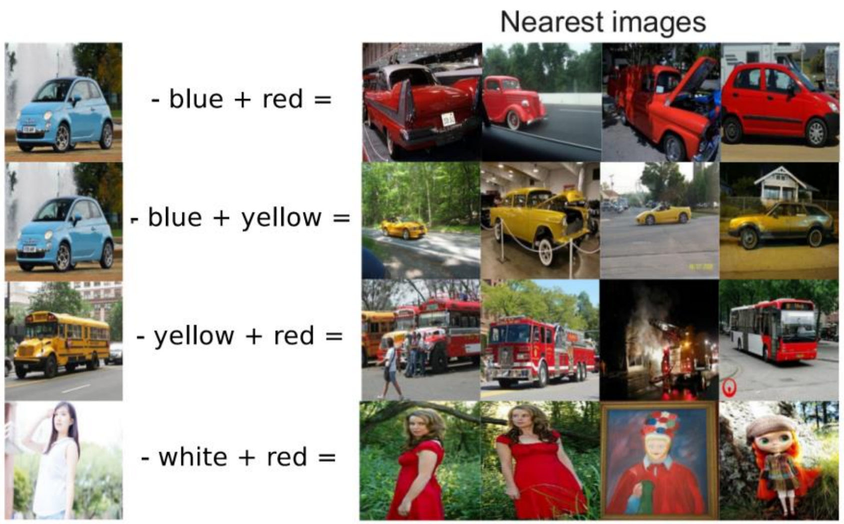
[Kim et al., ICASSP 2013]

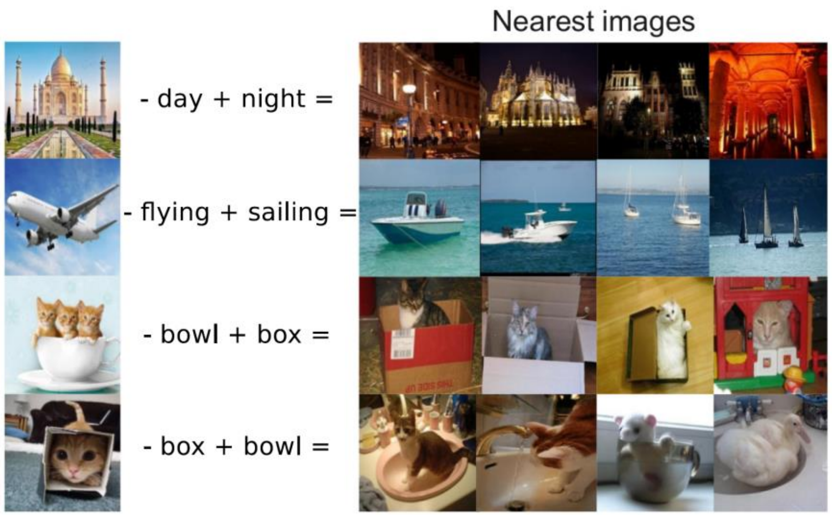
- Deep Boltzmann Machine



Core Challenge 3: Early Examples

Multimodal Vector Space Arithmetic



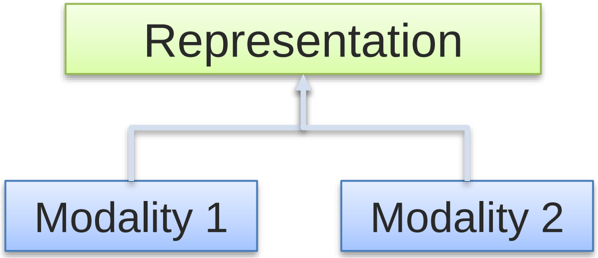


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

Core Challenge 3: Representation

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

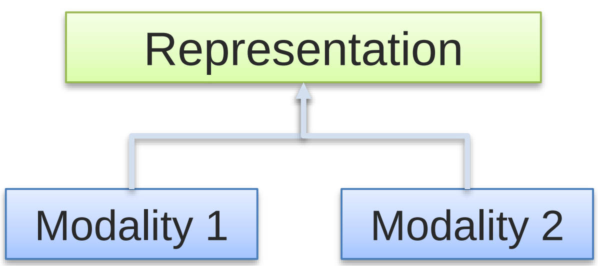
Ⓐ Joint representations:



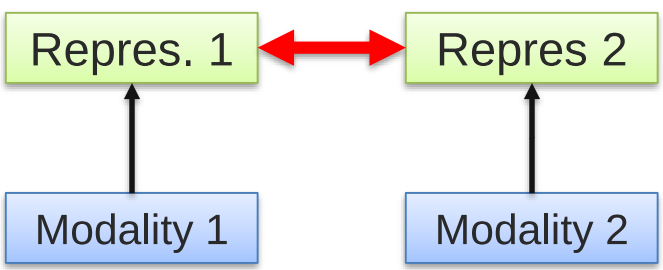
Core Challenge 3: Representation

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

A Joint representations:

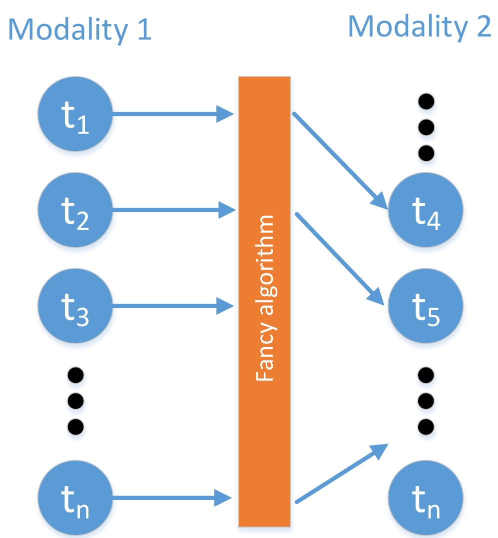


B Coordinated representations:



Core Challenge 4: Alignment

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



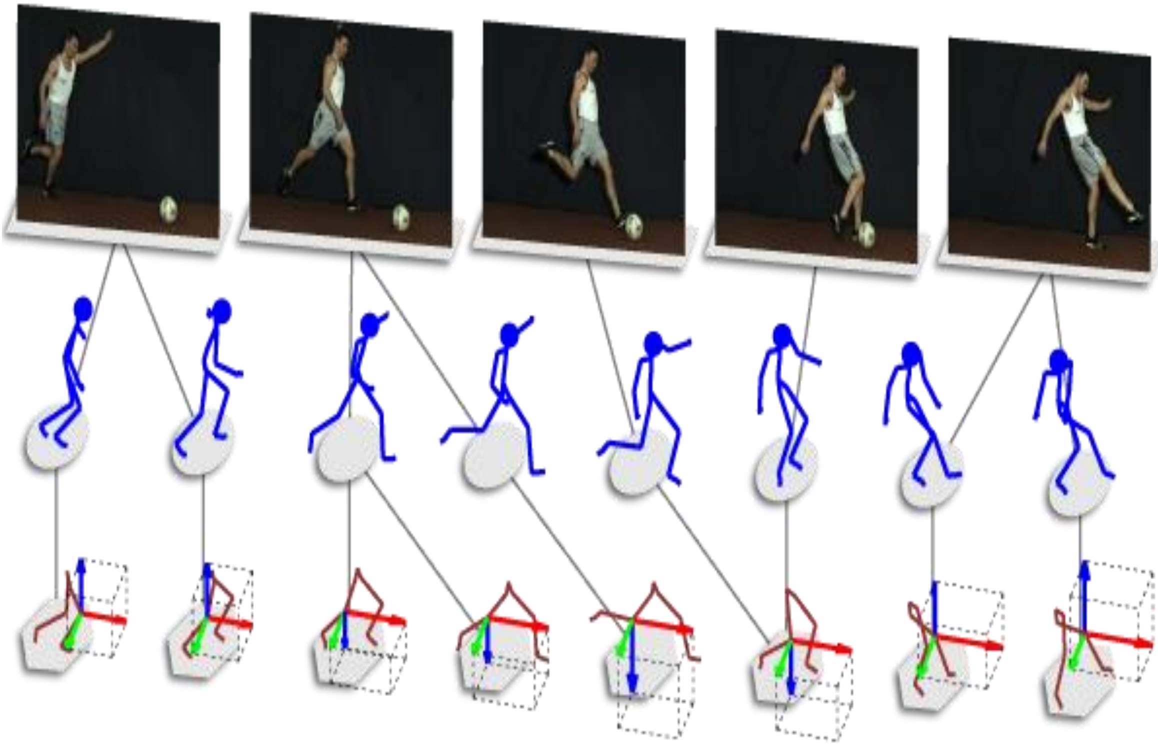
A **Explicit Alignment**

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

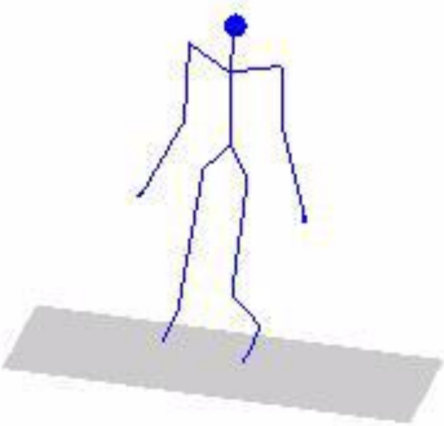
Core Challenge 4: Explicit Alignment



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

Core Challenge 4: Explicit Alignment

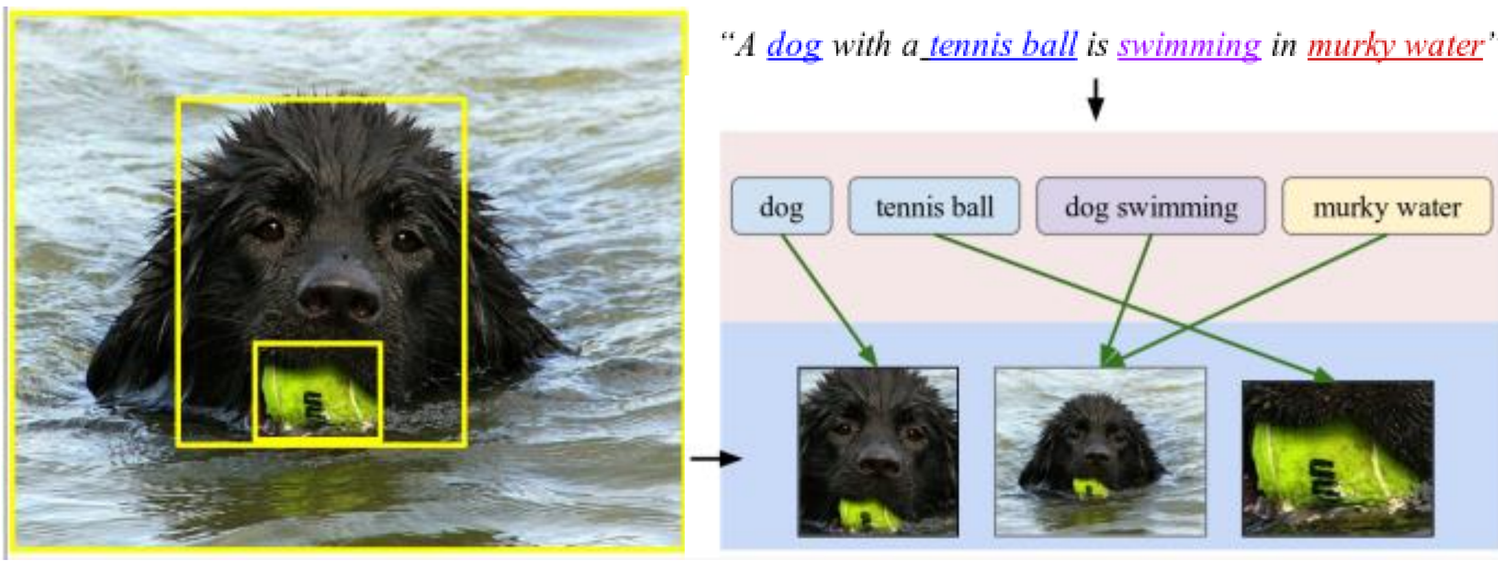
1/273



1/51

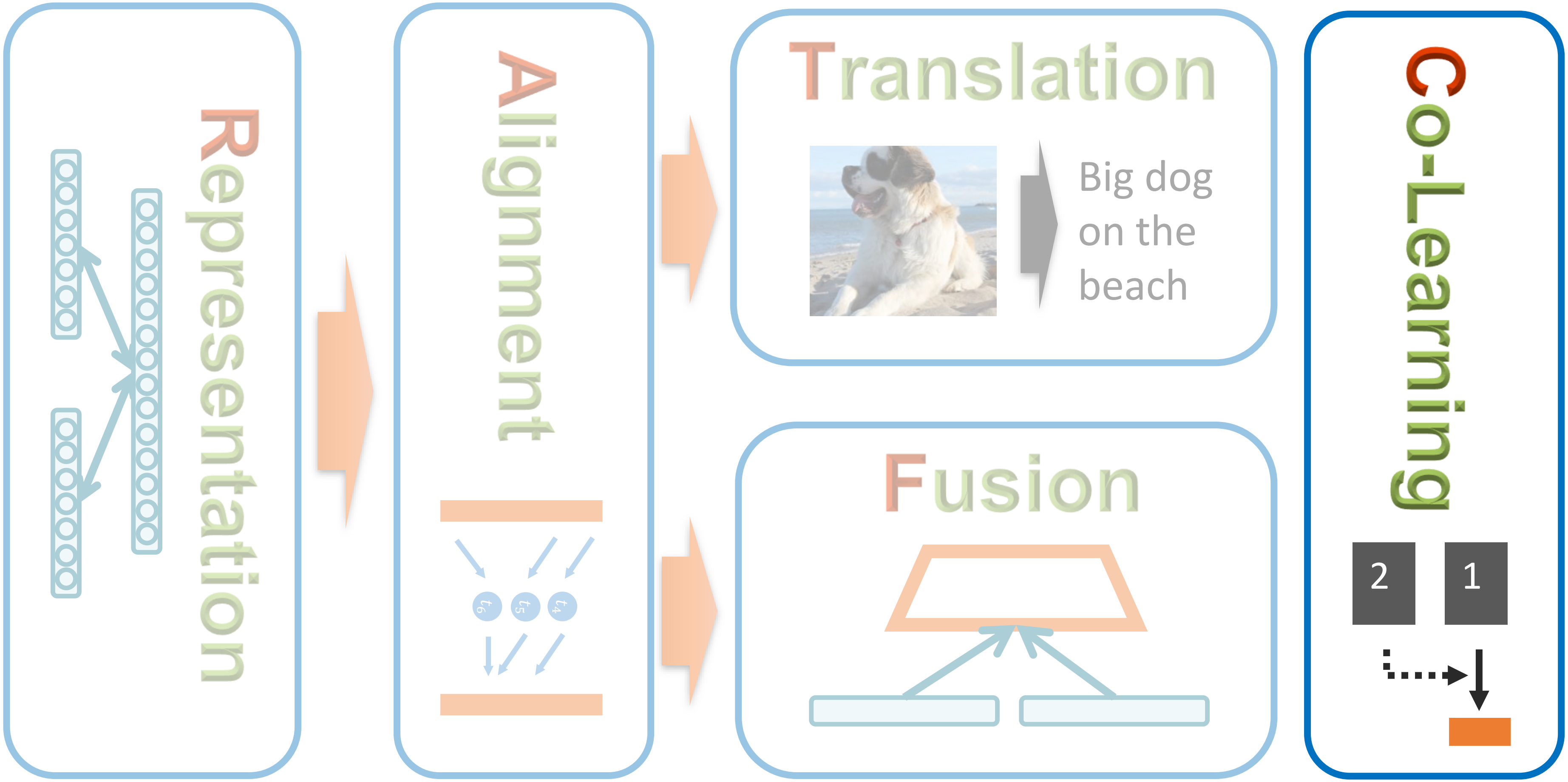
1/127

Core Challenge 4: Implicit Alignment

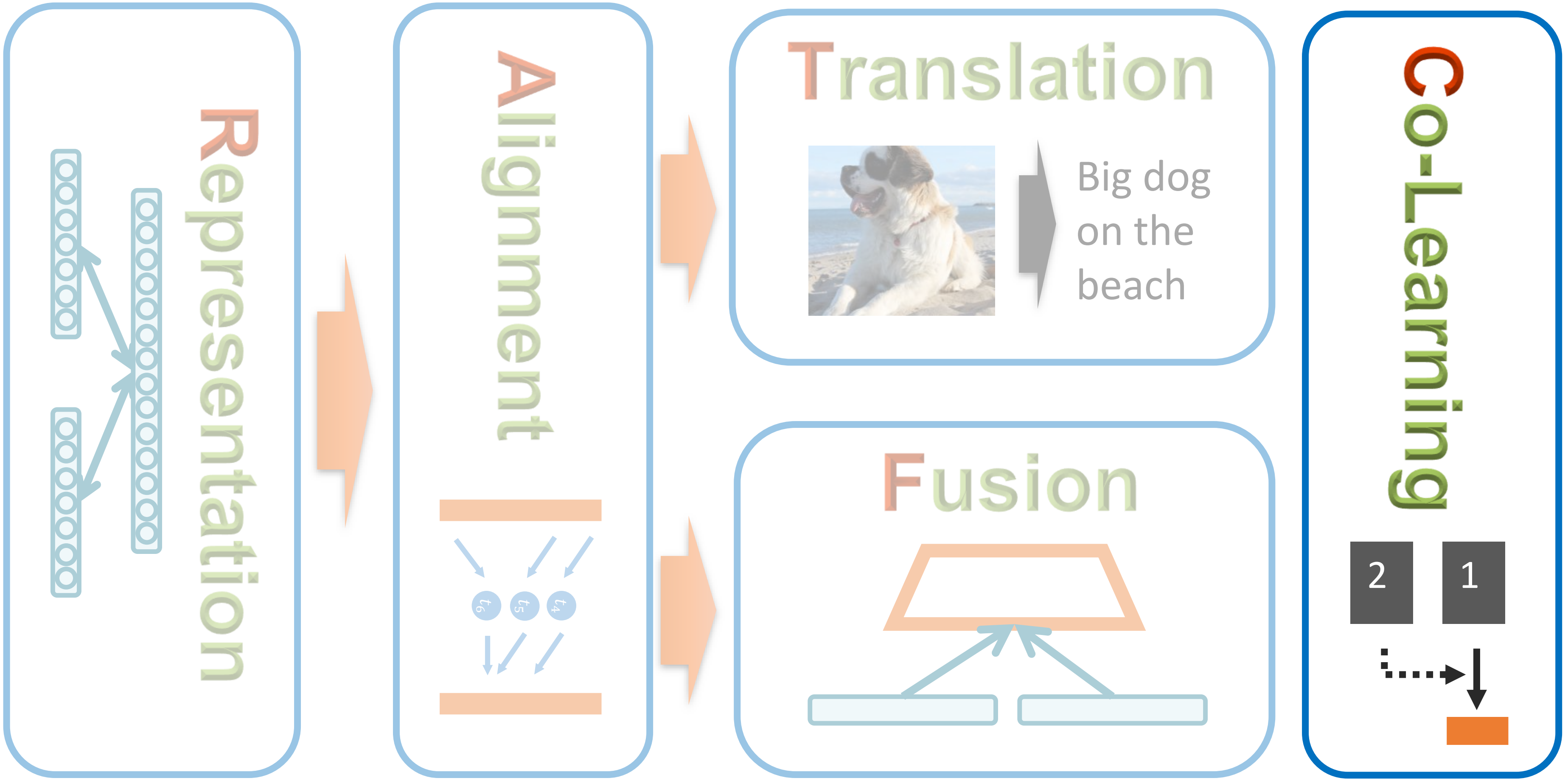


Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, https://arxiv.org/pdf/1406.5679.pdf

One Last Core Challenge

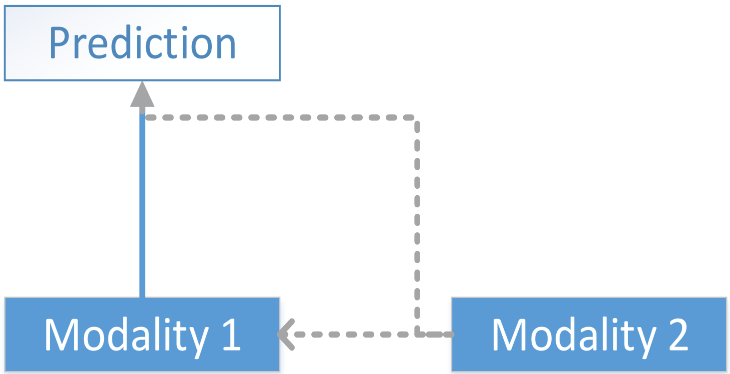


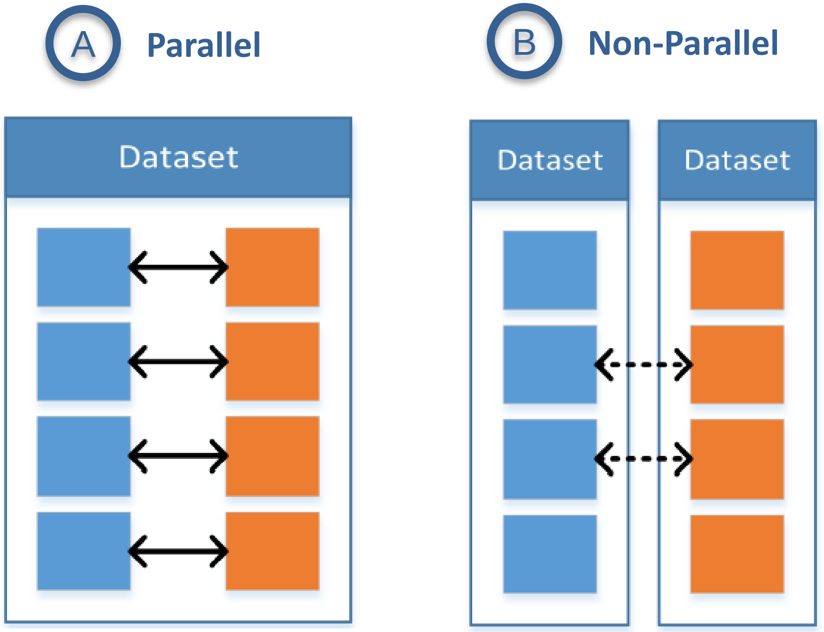
One Last Core Challenge



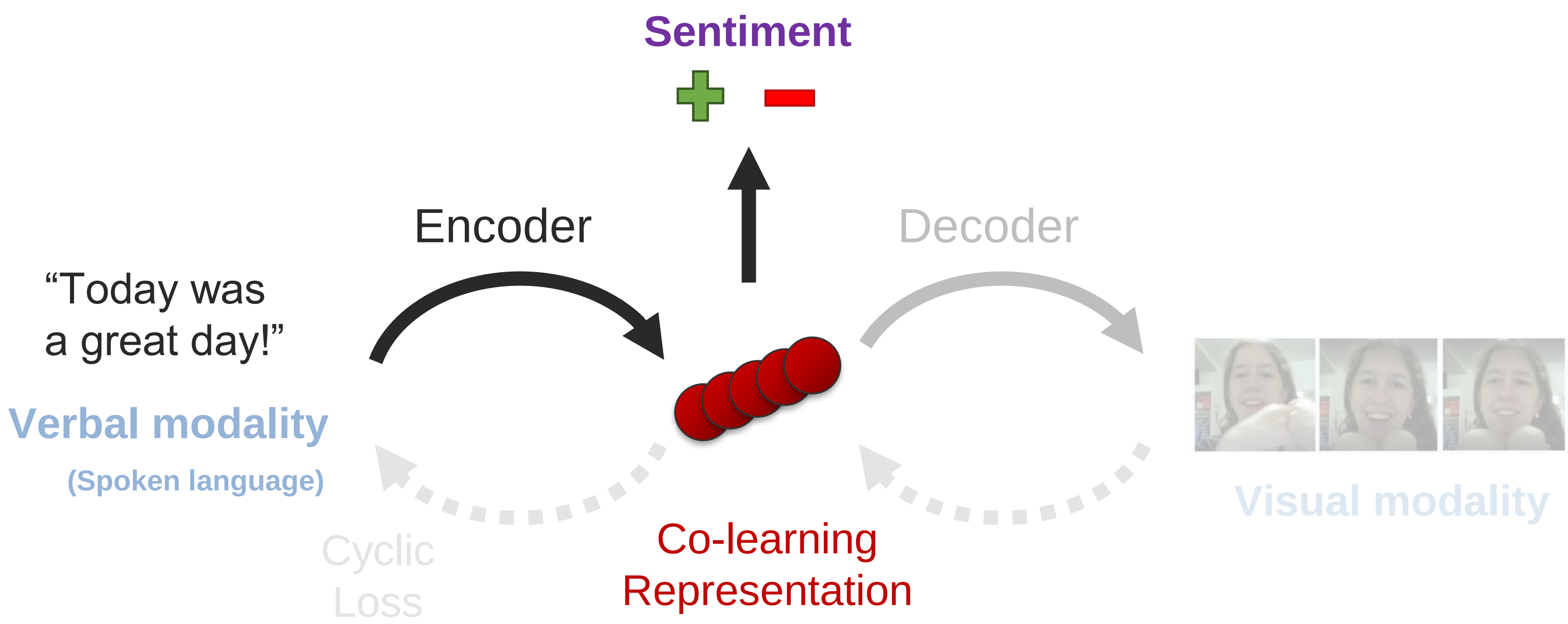
Core Challenge 5: Co-Learning

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



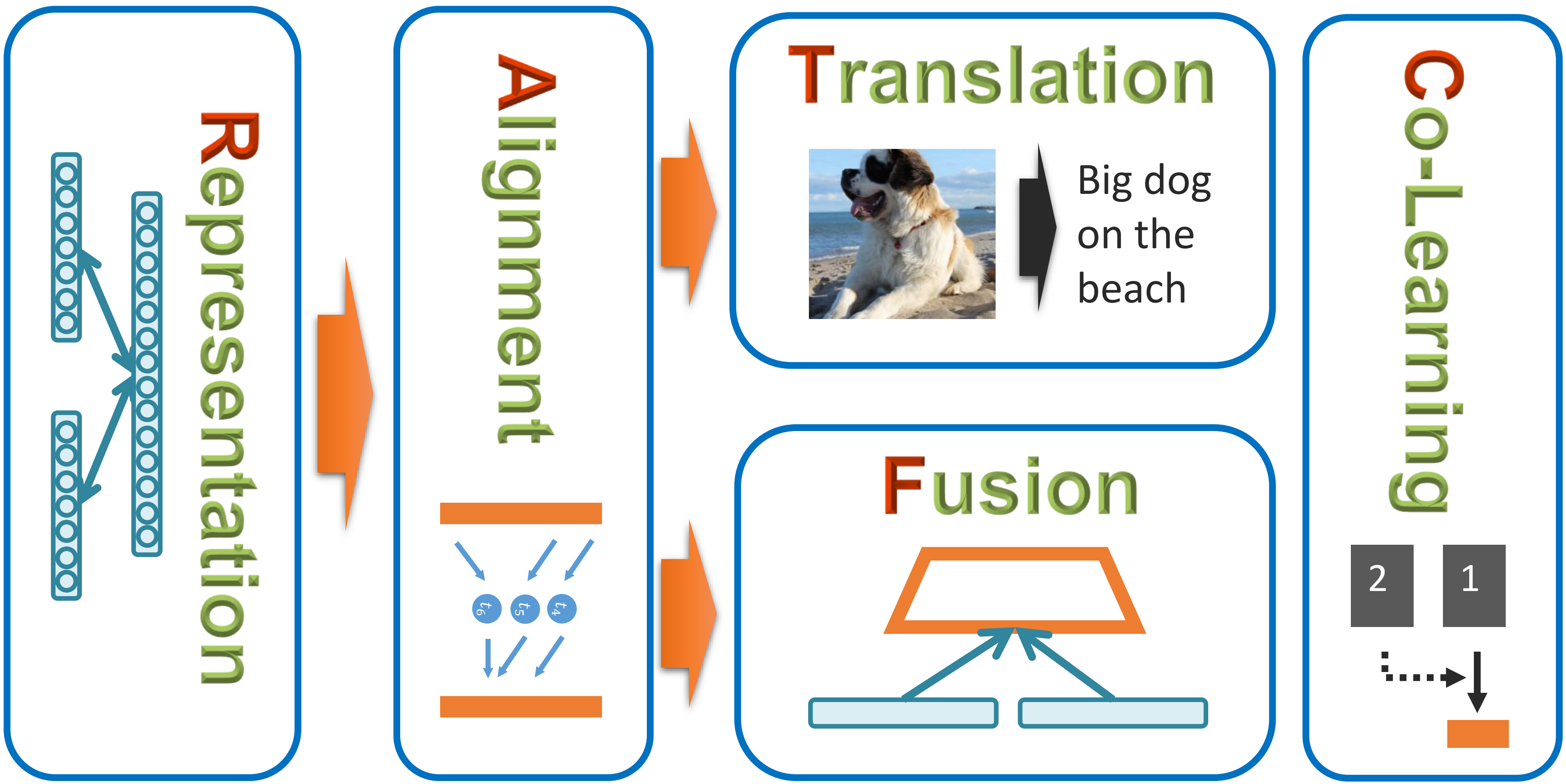


Core Challenge 5: Co-Learning



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

Multimodal Applications

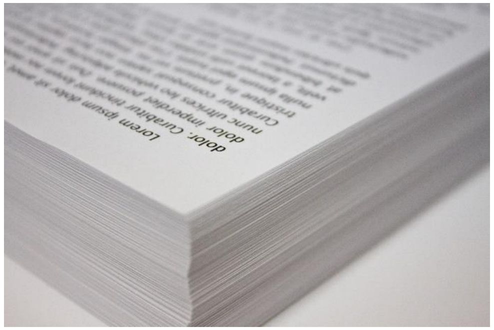
[https://arxiv.org/abs/1705.09406]

	Challenges				
Applications	Representation	Translation	Fusion	Alignment	Co-learning
Speech Recognition and Synthesis					
Audio-visual Speech Recognition	✓		✓	✓	✓
(Visual) Speech Synthesis	✓	✓			
Event Detection					
Action Classification	✓		✓		✓
Multimedia Event Detection	✓		✓		✓
Emotion and Affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media Description					
Image Description	✓	✓		✓	✓
Video Description	✓	✓	✓	✓	✓
Visual Question-Answering	✓		✓	✓	✓
Media Summarization	✓	✓	✓		
Multimedia Retrieval					
Cross Modal retrieval	✓	✓		✓	✓
Cross Modal hashing	✓				✓

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

Course Syllabus

Three Course Learning Paradigms



Reading assignments and course participation
(30% 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 lectures
(including guest lectures)

Course Recommendations and Requirements

1. Ready to read about 20 papers this semester !
   - Research papers as part of the weekly reading assignments
   - Asked to answer research questions about these papers
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
   - Three course project assignments
   - Designed to enhance state-of-the-art algorithms

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

Course Project

- Pre-proposal (in 2 weeks)
  - Define your dataset, research task and teammates
- First project assignment (in 5 weeks)
  - Experiment with unimodal representations
  - Explore/discuss simple baseline model(s)
- Midterm project assignment (in 10 weeks)
  - Implement and evaluate state-of-the-art model(s)
  - Discuss new multimodal model(s)
- Final project assignment (in 14 weeks)
  - 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 or 4 students
- The project should explore algorithmic novelty
- Possible venues for your final report:
  - NAACL 2020, ACL 2020, IJCAI 2020, ICML 2020
- We will discuss on Thursday about project ideas
- GPU resources available:
  - Amazon AWS and Google Cloud Platform

Examples of Previous Course Projects

- Select-Additive Learning: Improving Generalization in Multimodal Sentiment Analysis
  - https://arxiv.org/abs/1609.05244
- Preserving Intermediate Objectives: One Simple Trick to Improve Learning for Hierarchical Models
  - https://arxiv.org/abs/1706.07867
- Gated-Attention Architectures for Task-Oriented Language Grounding
  - https://arxiv.org/abs/1706.07230
- Efficient Low-rank Multimodal Fusion with Modality-Specific Factors
  - https://arxiv.org/abs/1806.00064

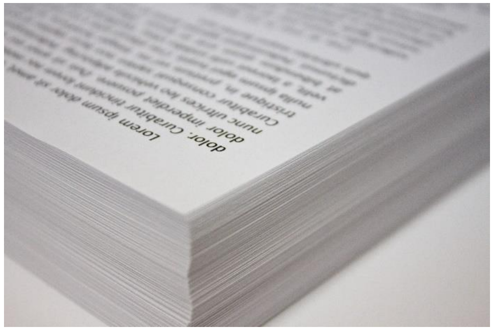
Examples of Previous Course Projects

- Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning
  - https://arxiv.org/abs/1802.00924
- Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities
  - https://arxiv.org/abs/1812.07809
- Self-Supervised Visual Representations for Cross-Modal Retrieval
  - https://arxiv.org/abs/1902.00378
- Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models
  - https://nips2018vigil.github.io/static/papers/accepted/33.pdf

Process for Selecting your Course Project

- **Thursday 8/29:** Lecture describing available multimodal datasets and research topics
- **Monday 9/2:** Submit a short paragraph listing your top 3 choices
- **Tuesday 9/3:** During the later part of the lecture, we will have a discussion period to help with team formation
- **Wednesday 9/11:** Pre-proposals are due. You should have selected your teammates, dataset and task

Course Grades



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

- Reading assignments 20%
- Lecture and course participation 10%

- Project preferences and proposal 5%
- First project assignment
  - Report and presentation 15%
- Mid-term project assignment
  - Report and presentation 20%
- Final project assignment
  - Report and presentation 30%

# 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 8/27 & 8/29	**Course introduction** • Research and technical challenges • Course syllabus and requirements	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 9/3 & 9/5 ***	**Basic concepts: neural networks** • Language, visual and acoustic • Loss functions and neural networks	**Basic concep** • Gradient • Practical
Week 3 9/10 & 9/12 **Pre-proposal**	**Convolutional neural networks** • Convolutional kernels and CNNs • Residual networks	**Recurrent ne** • Gated ne • Backprop
Week 4 9/17 & 9/19	**Multimodal representation learning** • Multimodal auto-encoders • Multimodal joint representations	**Coordinated representations** • Deep canonical correlation analysis • Non-negative matrix factorization
Week 5 9/24 & 9/26	**Modular and factorized representations** • Module networks • Factorized representations	**Unsupervised representations** • Variational auto-encoder • Generative adversarial approaches
Week 6 10/1 & 10/3	***First project assignment - Presentations***	

Project preferences due on Monday night

Pre-proposals due on Sunday 9/11

First assignment due on Sunday 10/6

Lecture Schedule

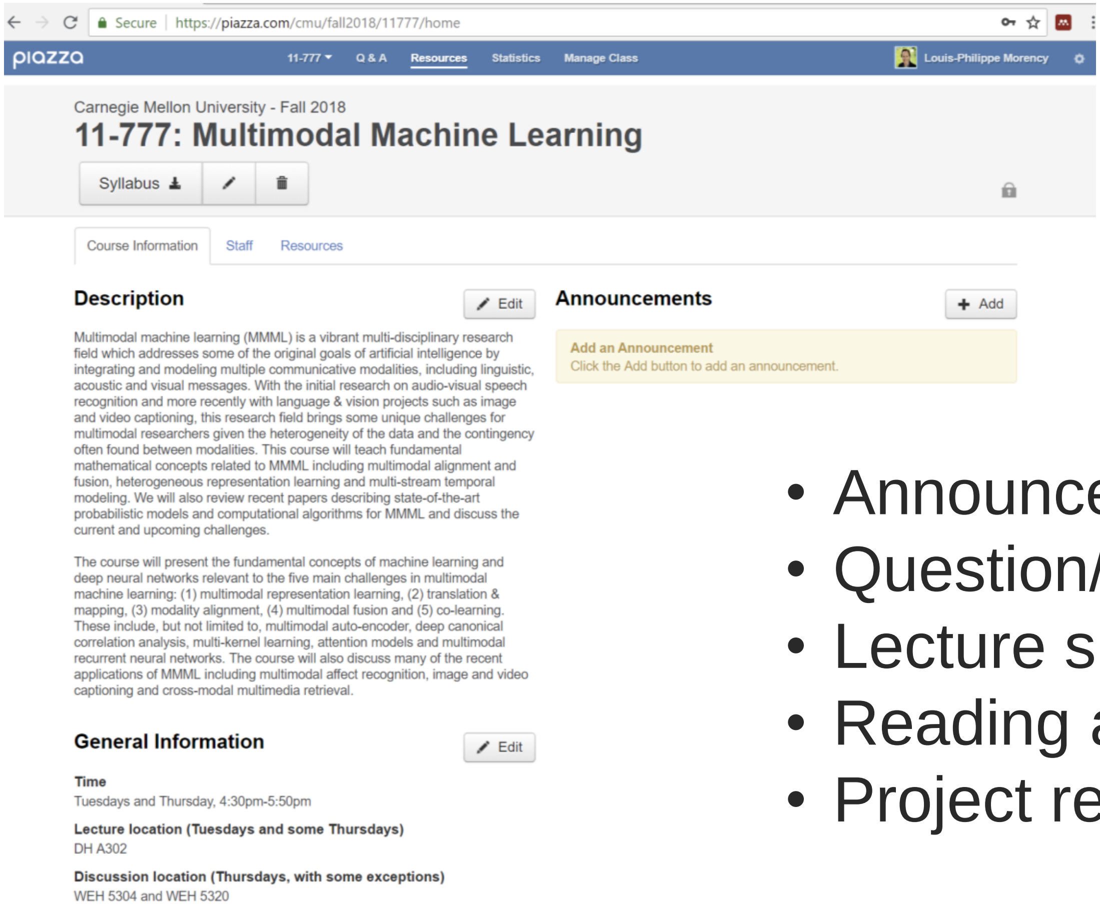
Classes	Tuesday Lectures	Thursday Lectures
Week 7 10/8 & 10/10	**Multimodal alignment** • Explicit - dynamic time warping • Implicit - attention models	**Alignment and representation** • Multi-head attention • Multimodal transformers
Week 8 10/15 & 10/17 ***	**Reinforcement learning** • Markov decision process • Q learning and policy gradients	**Multimodal RL** • Deep Q learning • Multimodal applications
Week 9 10/22 & 10/24 ***	**Probabilistic graphical models** • Dynamic Bayesian networks • Coupled and factor HMMs	**Discriminative graphical models** • Boltzmann distribution and CRFs • Continuous and fully-connected CRFs
Week 10 10/29 & 10/31	**Multimodal fusion and co-learning** • Multi-kernel learning and fusion • Multimodal transfer learning	**New directions in Multimodal ML** • Overview of recent approaches in multimodal machine learning
Week 11 11/5 & 11/7	***Mid-term project assignment - Presentations***	

Midterm due on 11/10.

Lecture Schedule

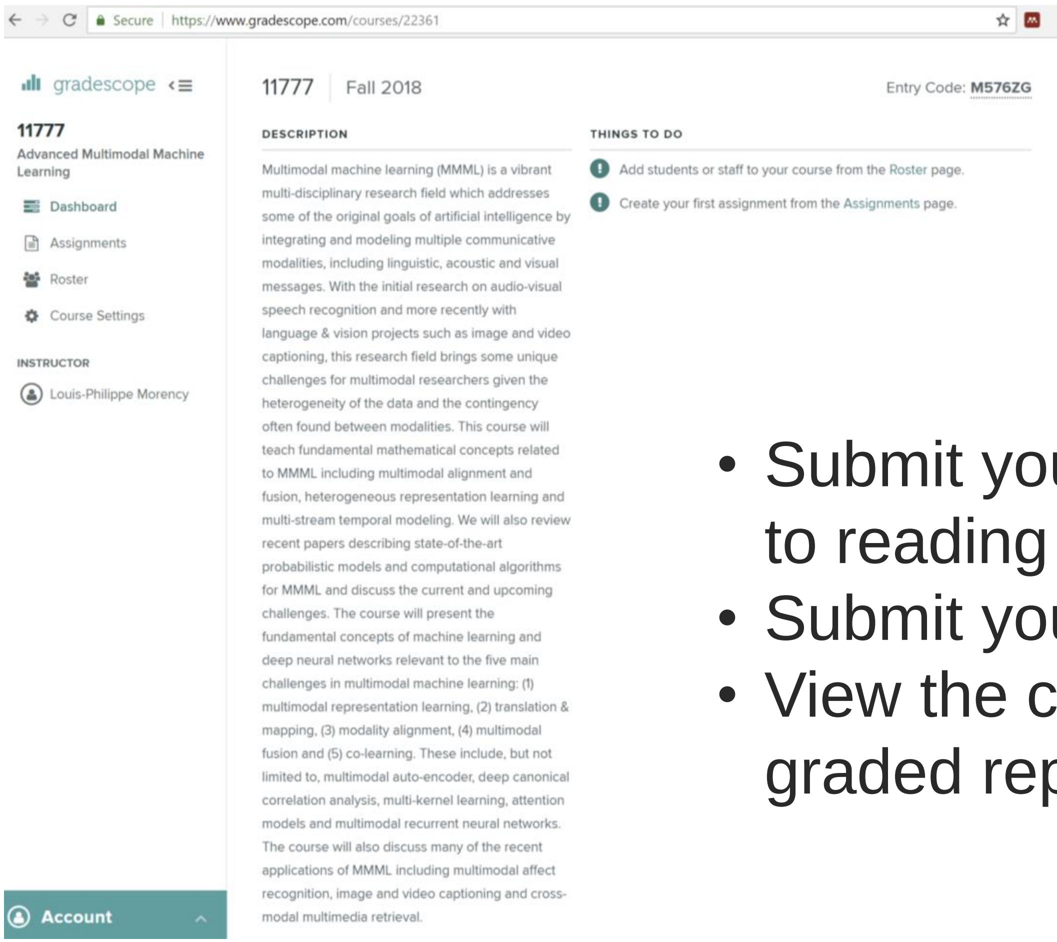
Classes	Tuesday Lectures	Thursday Lectures
Week 12 11/12 & 11/14	**Multi-lingual representations** • Neural machine translation • Guest lecture: Graham Neubig	**Knowledge representation** ● Multimodal knowledge discovery ● Guest lecture
Week 13 11/19 & 11/21	***Thanksgiving week (no classes)***	
Week 14 11/26 & 11/28	**Language, vision and action** • Neural machine translation • Guest lecture	**Multimodal affective computing** • Emotion and sentiment analysis • Guest lecture
Week 15 12/3 & 12/5	***Final project assignment - Presentations***	Final due on 12/8.

Piazza https://piazza.com/cmu/fall2019/11777/home



- Announcements
- Question/Answers
- Lecture slides
- Reading assignments
- Project resources

Gradescope – Entry Code: 9NGKXX



- Submit your weekly answers to reading questions
- Submit your project reports
- View the comments from your graded reports

Project Preferences – Due Monday 9/3

- 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/fall2019/11777/home