



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 1.2: Multimodal Research Tasks

Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk. Some slides from Graham Neubig.*

Lecture Objectives

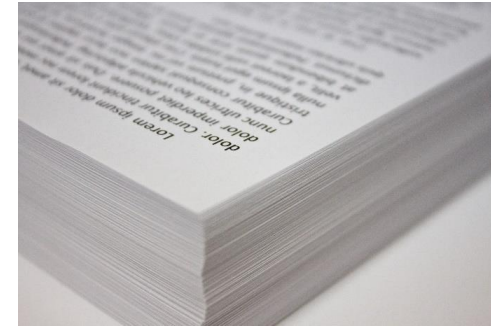
- Course syllabus and project assignments
 - Course and assignment schedule
 - Projects and team matching
 - Grades and course structure
- A historical view on multimodal research
- Experimental design
 - Research questions and hypotheses
- Multimodal datasets and research tasks
 - 100+ multimodal datasets (+ curated list)
- Examples of previous course projects

Course Syllabus

Three Course Learning Paradigms



Course lecture participation
(16% of your grade)




Reading assignments
(12% of your grade)





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

Course project assignments
(72% of your grade)

Piazza <https://piazza.com/cmu/fall2022/11777/info>

PIAZZA 11777 Q & A Resources Statistics Manage Class  Louis-Philippe Morency

Carnegie Mellon University - Fall 2022
11777: Multimodal Machine Learning

Syllabus    

Course Information Staff Resources

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. This course will teach fundamental mathematical concepts related to MMML including multimodal alignment and fusion, heterogeneous representation learning and multi-stream 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.

Recommended preparation: 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. Students should have proper academic background in probability, statistic and linear algebra. Programming knowledge in Python is also strongly recommended.

More details in the Syllabus document.

General Information

Time
Tuesdays and Thursday, 10:10am-11:30am

Location
DH 1212

Announcements

Add an Announcement
Click the Add button to add an announcement.

- ✓ Announcements
- ✓ Question/Answers
- ✓ Reading assignments
- ✓ Project resources
- ✓ Course syllabus

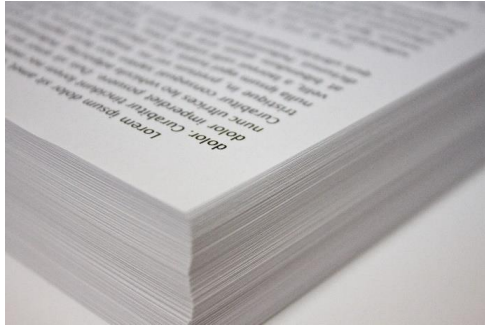
Copyright © 2022 Piazza Technologies, Inc. All Rights Reserved.



Course Recommendations and Requirements

- 1 Ready to read about 6 papers this semester !**
 - Curated list of research papers for the 6 reading assignments
 - Summarize one paper and contrast it with other 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**
 - Projects are designed to enhance state-of-the-art algorithms
 - Four project assignments, to help scaffold the project tasks

Course Grades



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

- Lecture highlights 16%
- Reading assignments 12%

- Project preferences/pre-proposal 2%
- First project assignment 10%
- Second project assignment 10%
- Mid-term project assignment
 - Report and presentation 20%
- Final project assignment
 - Report and presentation 30%

Lecture Highlight Form

Lecture 2.1 - Highlight Form

DEADLINE Submit your Lecture Highlight form by Thursday Sept 10, 2020 at 10:40am EST. You have 42 hours to fill out this form, from the scheduled end time of the lecture.

IMPORTANT: Please read the detailed instructions in Piazza's Resources section ("Lecture Highlights - Instructions.pdf", in the Instructions for Course Assignments list) before filling out this form.

<https://piazza.com/cmu/fall2020/11777a/resources>

Your email address (**Imorency@andrew.cmu.edu**) will be recorded when you submit this form. Not you? [Switch account](#)

* Required

First 30 mins - Main take home message (about 15-40 words) * 2 points

Your answer

(Optional) First 30 mins - Any question? Please include slide number(s)

Your answer

Next 30 mins - Main take home message (about 15-40 mins) * 2 points

Your answer

Similar to note-taking during lectures

- ➡ For each course segment (30mins):
2 sentences describing the main points

Help you summarizing the lecture

- ➡ What is the main take-away message from the lecture
Short paragraph (15-40 words)

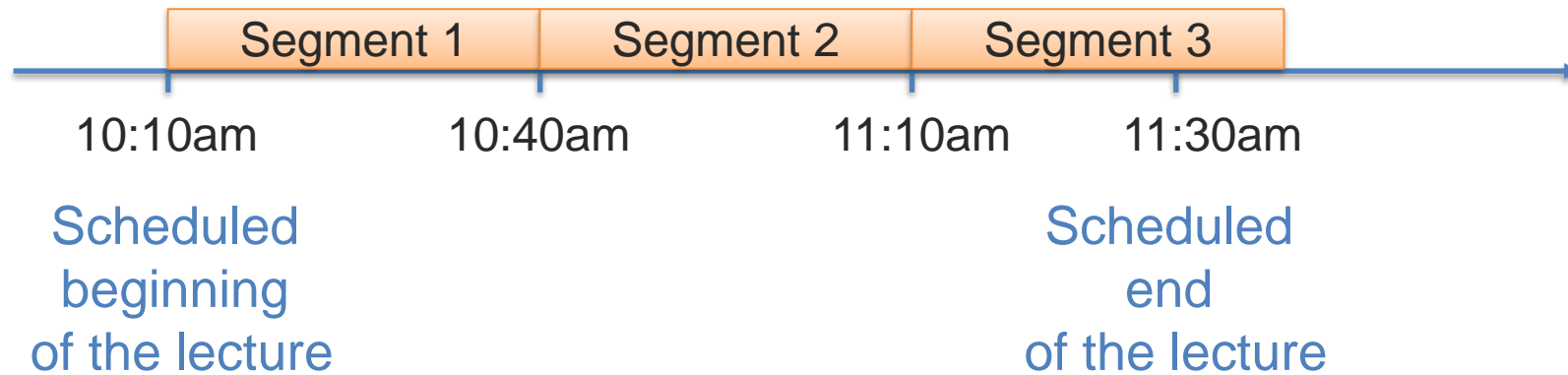
Ask questions about the lecture

- ➡ Will be answered either online or at the next lecture

Submitted same day as lecture (before 11:59pm)

- ➡ Students are encouraged to attend lectures in person

Lecture Highlight Form - Segments



- ➔ Segment 1 starts at 10:10am, even if the lecture starts slightly later.
- ➔ Segment 3 ends whenever the lecture ends
- ➔ Slides happening around the segment borders (± 5 min of 10:40am and 11:10am) can be included in either neighboring segment.

First Reading Assignment – Week 2

- Study groups: 9-10 students per group (randomly, in Piazza)
- 3 paper options are available
 - **Each student should pick one paper option!**
 - Google Sheets were created to help balance the papers between group members
 - Then you will create a short summary to help others [1 point]
- Discussions with your study group
 - Read other's summaries. Ask questions!
 - Write follow-up posts comparing the papers and suggesting ideas [1 point]
 - At least one follow-up post for every paper you did not read



First Reading Assignment – Week 2

Four main steps for the reading assignments

1. **Monday 8pm:** Official start of the assignment
2. **Wednesday 8pm:** Select your paper
3. **Friday 8pm:** Post your summary
4. **Monday 8pm:** Post your follow-up posts

Detailed instructions posted on Piazza

<https://piazza.com/cmu/fall2022/11777/resources>



Late Submissions and Wildcards

- Each student has **6** late submission wildcards
 - For lecture highlight forms or reading assignments
- Each project team has **2** late submission wildcards
 - For any of the project assignments
- Total number of wildcards: 8 (6 individual and 2 team-level)
- Each wildcard gives 24-hour extension
 - No partial credits for the wildcards
 - Automatically calculated (no need to contact us apriori)

See details about late submission policy in syllabus

<https://piazza.com/cmu/fall2022/11777/resources>



Course Project Guidelines

- Preference for datasets with at least these two modalities:
 - Natural language and visual/images
- Teams of 3, 4 or 5 students
- The project should explore new research ideas
- Possible venues for your final report:
 - ACL 2023, IJCAI 2023, ICML 2023, ICMI 2023
- GPU resources available (more details soon!):
 - Amazon AWS and Google Cloud Platform

Course Project Timeline

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

Pre-proposal (*due Wednesday Sept. 14*)

- Define your dataset, research task and teammates

First project assignment (*due Sunday Sept. 25*)

- Study related work to your selected research topic

Second project assignment (*due Sunday Oct 9*)

- Experiment with unimodal representations

Midterm project assignment (*due Sunday Oct 30*)

- Implement and evaluate state-of-the-art model(s)

Final project assignment (*due Sunday Dec. 11*)

- Implement and evaluate new research ideas

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

Process for Selecting your Course Project

- **Thursday 9/1 (today!):** multimodal dataset lecture
- **Tuesday 9/6:** Let us know your dataset preferences
 - Preferences will be shared with all students
- **Thursday 9/8:** During the later part of the lecture, we will have an interactive period to help with team formation. More details to come
- **Wednesday 9/14:** Pre-proposals are due. You should have selected your teammates, dataset and task

 **Special office hours by TAs** (schedule will be posted on Piazza)

Project Preferences – Due Tuesday 9/6

- Post your project preferences:
 - Select your preferred research areas
 - List preferred multimodal datasets
 - Use alphanumeric code of each dataset
 - Detailed dataset list on Piazza (including slides in the Lecture 1.2 appendix)
 - Examples of research ideas you are interested in
 - For topics or datasets not listed:
 - Include a description with links (for other students)

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures	
Week 1 8/30 & 9/1	Course introduction <ul style="list-style-type: none"> Multimodal core challenges Course syllabus 	Multimodal applications and datasets <ul style="list-style-type: none"> Research tasks and datasets Team projects 	
Week 2 9/6 & 9/8 <i>Read due: 9/9</i>	Basic concepts: neural networks <ul style="list-style-type: none"> Loss functions and neural networks Gradient and optimization 	Unimodal representations <ul style="list-style-type: none"> Dimensions of heterogeneity Visual representations 	<div style="border: 2px solid red; padding: 5px; text-align: center;"> Project preferences due on Tuesday 9/6 </div>
Week 3 9/13 & 9/15 <i>Read due: 9/16</i> <i>Proj. Due: 9/14</i>	Unimodal representations <ul style="list-style-type: none"> Language representations Signals, graphs and other modalities 	Multimodal representations <ul style="list-style-type: none"> Cross-modal interactions Multimodal fusion 	<div style="border: 2px solid red; padding: 5px; text-align: center;"> Pre-proposals due on Wednesday 9/14 </div>
Week 4 9/20 & 9/22 <i>Proj. due: 9/25</i>	Multimodal representations <ul style="list-style-type: none"> Coordinated representations Multimodal fission 	Multimodal alignment and grounding <ul style="list-style-type: none"> Explicit alignment Multimodal grounding 	<div style="border: 2px solid red; padding: 5px; text-align: center;"> First assignment due on Sunday 9/25 </div>
Week 5 9/27 & 9/29 <i>Read due: 9/30</i>	<i>Project hours (Research ideas)</i>	Aligned representations <ul style="list-style-type: none"> Self-attention transformer models Masking and self-supervised learning 	
Week 6 10/4 & 10/6 <i>Proj. due: 10/9</i>	Multimodal aligned representations <ul style="list-style-type: none"> Multimodal transformers Video and graph representations 	Multimodal Reasoning <ul style="list-style-type: none"> Structured and hierarchical models Memory models 	<div style="border: 2px solid red; padding: 5px; text-align: center;"> Second assignment due on Sunday 10/9 </div>

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 7 10/11 & 10/13 Read due: 10/14	Multimodal Reasoning <ul style="list-style-type: none">Reinforcement learningDiscrete structure learning	Multimodal Reasoning <ul style="list-style-type: none">Logical and causal inferenceExternal knowledge
Week 8 10/18 & 10/20	Fall Break – No lectures	
Week 9 10/25 & 10/27 Proj. due: 10/30	Generation <ul style="list-style-type: none">Translation, summarization, creationGenerative models: VAEs	Generation <ul style="list-style-type: none">GANs and diffusion modelsModel evaluation and ethics
Week 10 11/1 & 11/3	Project presentations (midterm)	Project presentations (midterm)
Week 11 11/8 & 11/10 Read due: 11/12	Transference <ul style="list-style-type: none">Modality transferMultimodal co-learning	Quantification <ul style="list-style-type: none">Heterogeneity and interactionsBiases and fairness
Week 12 11/15 & 11/17 Read due: 11/21	Project hours (Research ideas)	New research directions <ul style="list-style-type: none">Recent approaches in multimodal ML

Midterm assignment
due on Sunday 10/30

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 13 11/22 & 11/24	Thanksgiving Week – No Class –	
Week 14 11/30 & 12/2	Language, Vision, and Actions <ul style="list-style-type: none">• Motion and navigation• Robots and embodied AI	Multimodal Applications <ul style="list-style-type: none">• Healthcare and affective computing• Artificial social intelligence
Week 15 12/6 & 12/8 <i>Proj. due: 12/11</i>	Project presentations (final)	Project presentations (final)

Final assignment due on Sunday 12/11

Spring 2023 Edition of the MMML Course !



Yonatan Bisk

ybisk@cs.cmu.edu

<https://yonatanbisk.com/>



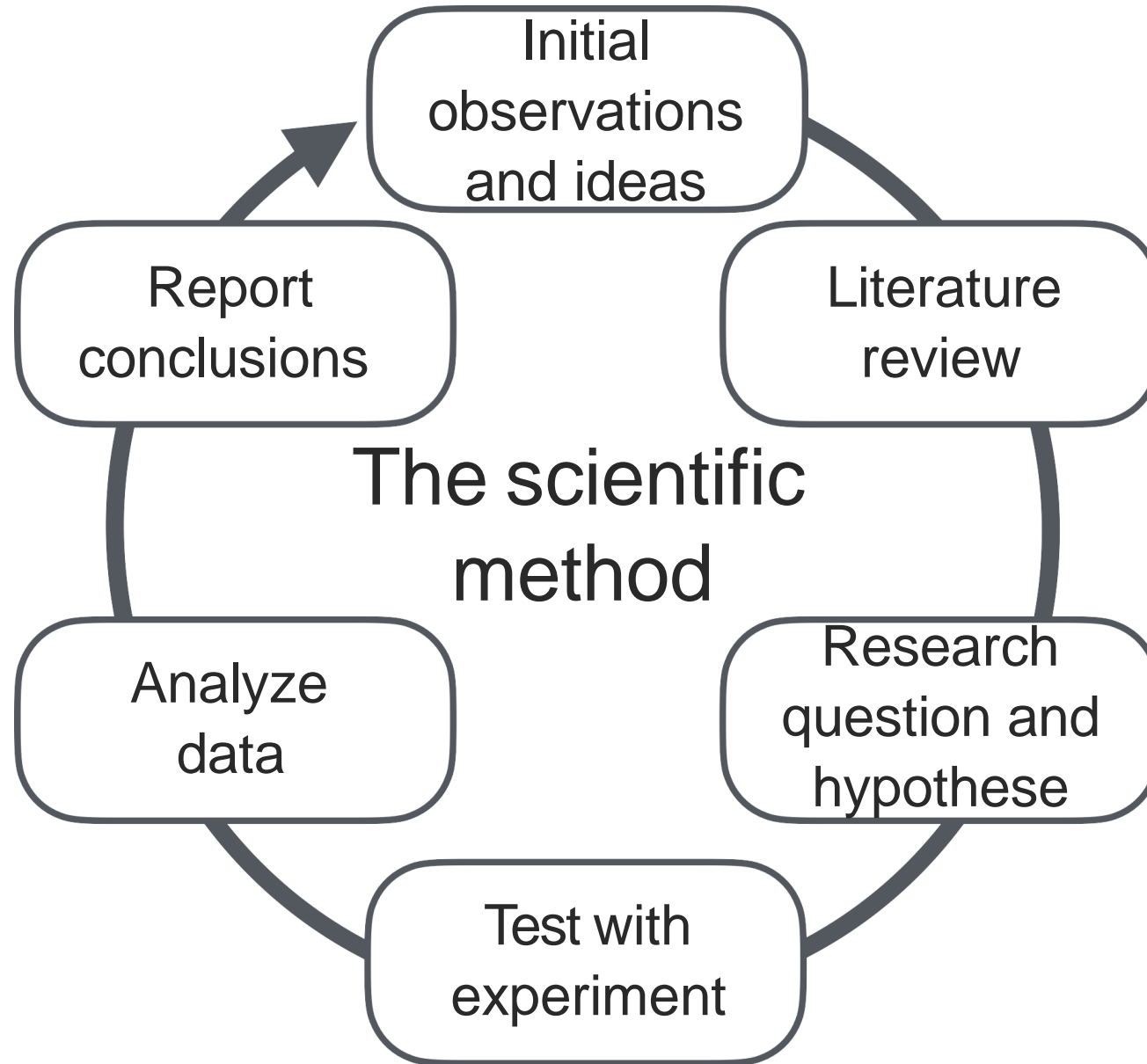
Daniel Fried

dfried@cs.cmu.edu

<https://dpfried.github.io/>

Experimental Design

(aka, finding a good research idea for your project)



Credit: Adapted From Wikipedia (Efbrasil)

How Do We Get Research Ideas?

Turn a concrete understanding of existing research's failings to a higher-level experimental question.

- **Bottom-up Discovery** of research ideas
- Great tool for incremental progress, but may preclude larger leaps

Move from a higher-level question to a lower-level concrete testing of that question.

- **Top-down Design** of research ideas
- Favors bigger ideas, but can be disconnected from reality

Bottom-Up Discovery

The 11-777 midterm project assignment will enable this bottom-up discovery:

1. Experiment state-of-the-art models
2. Analyze successes and failures of these models
3. Identify ways you could improve on these failure cases

Your research ideas will evolve during the semester!

Top-down Design

Brainstorming: Take the time to brainstorm with your teammates, with TAs and with instructors.

- Office hours with TAs these coming 2 weeks
- Project hours with instructors in the next month
- Communicate with us via Piazza!

Literature review: The first assignment will allow you to review recent work related to your dataset and your initial research ideas

- When exploring the dataset (second assignment), you should also expand your research ideas

Scientific Research Questions and Hypotheses

Research Questions

- One or several explicit questions regarding the thing that you want to know
- Hypotheses are easier to draft with “Yes-no” questions than “how to” questions

Hypothesis:

- What you think the answer to the question may be a-priori
- Should be *falsifiable*: if you get a certain result the hypothesis will be validated, otherwise disproved

Questions + Hypotheses

Are All Languages Equally Hard to Language-Model?

Modern natural language processing practitioners strive to create modeling techniques that work well on all of the world's languages. Indeed, most methods are portable in the following sense: Given appropriately annotated data, they should, in principle, be trainable on any language. However, despite this crude cross-linguistic compatibility, it is unlikely that all languages are equally easy, or that our methods are equally good at all languages.

Cotterell et al. (2018)

What makes a particular podcast broadly engaging?

As a media form, podcasting is new enough that such questions are only beginning to be understood (Jones et al., 2021). Websites exist with advice on podcast production, including language-related tips such as reducing filler words and disfluencies, or incorporating emotion, but there has been little quantitative research into how aspects of language usage contribute to listener engagement.

Reddy et al. (2018)

Beware "Does X Make Y Better?" "Yes"

The above question/hypothesis is natural, but indirect

- If the answer is "no" after your experiments, how do you tell what's going wrong?

Usually you have an intuition about *why* X will make Y better (not just random)

Can you think of other research questions/ hypotheses that confirm/falsify these assumptions

Exploratory Research Questions

- These questions will be more open-ended
- This is a valid part of research, but you have to be careful about your conclusion claims

For the course research project, exploratory questions are also good options

Examples of Research Ideas

~~State of the art prediction performance on dataset XYZ~~

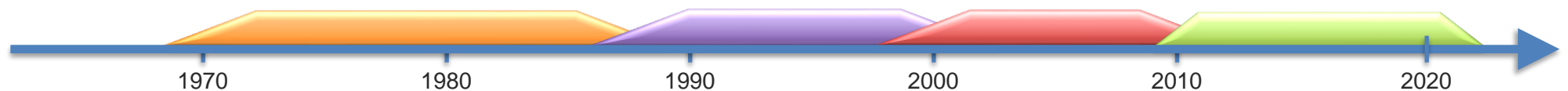
- Robustness to missing/noisy modalities, adversarial attacks
- Studying social biases and creating fairer models
- Interpretable and trustworthy models
- Understanding compositionality and multimodal reasoning
- Better understanding of the cross-modal interactions in multimodal models
- Faster and more efficient models for training, storage and inference
- Theoretical projects are welcome too
 - Make sure that you have experiments to validate and test your theory

Multimodal Research: A Historical View

Prior Research in “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



Behavioral Study of Multimodal



Language
and gestures

David McNeill

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

McGurk effect



Behavioral Study of Multimodal



Language
and gestures

David McNeill

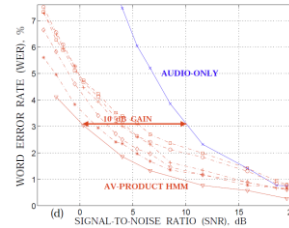
“For McNeill, gestures are in effect the speaker’s thought in action, and integral components of speech, not merely accompaniments or additions.”

McGurk effect



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

1) Audio-Visual Speech Recognition



Redundancy between audio and visual modalities help with handling noise and with robustness

2) Multimodal interfaces



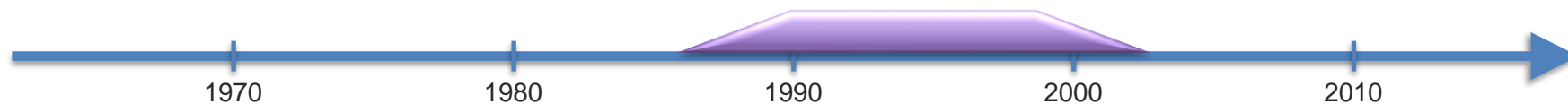
Affective Computing is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena.

3) Multimedia



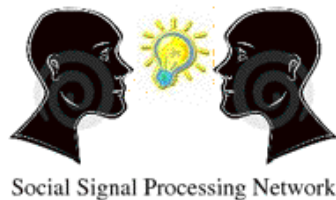
[1994-2010]

“...automatically combines speech, image and natural language understanding to create a full-content searchable digital video library.”



➤ The “Interaction” Era (2000s)

Modeling Multimodal Social Interactions



AMI Project [2001-2006, IDIAP]

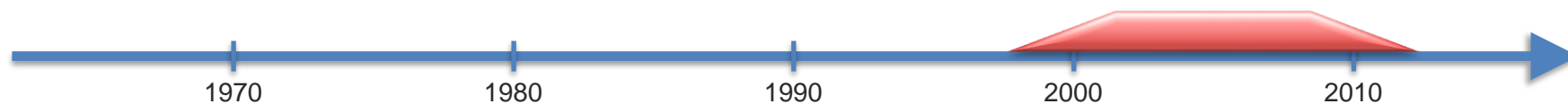
- 100+ hours of meeting recordings
- Transcribed and annotated

CALO Project [2003-2008, SRI]

- Cognitive Assistant that Learns and Organizes
- Siri was a spinoff from this project

SSP Project [2008-2011, IDIAP]

- Social Signal Processing
- Great dataset repository: <http://sspnet.eu/>



➤ The “deep learning” era (2010s until ...)

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

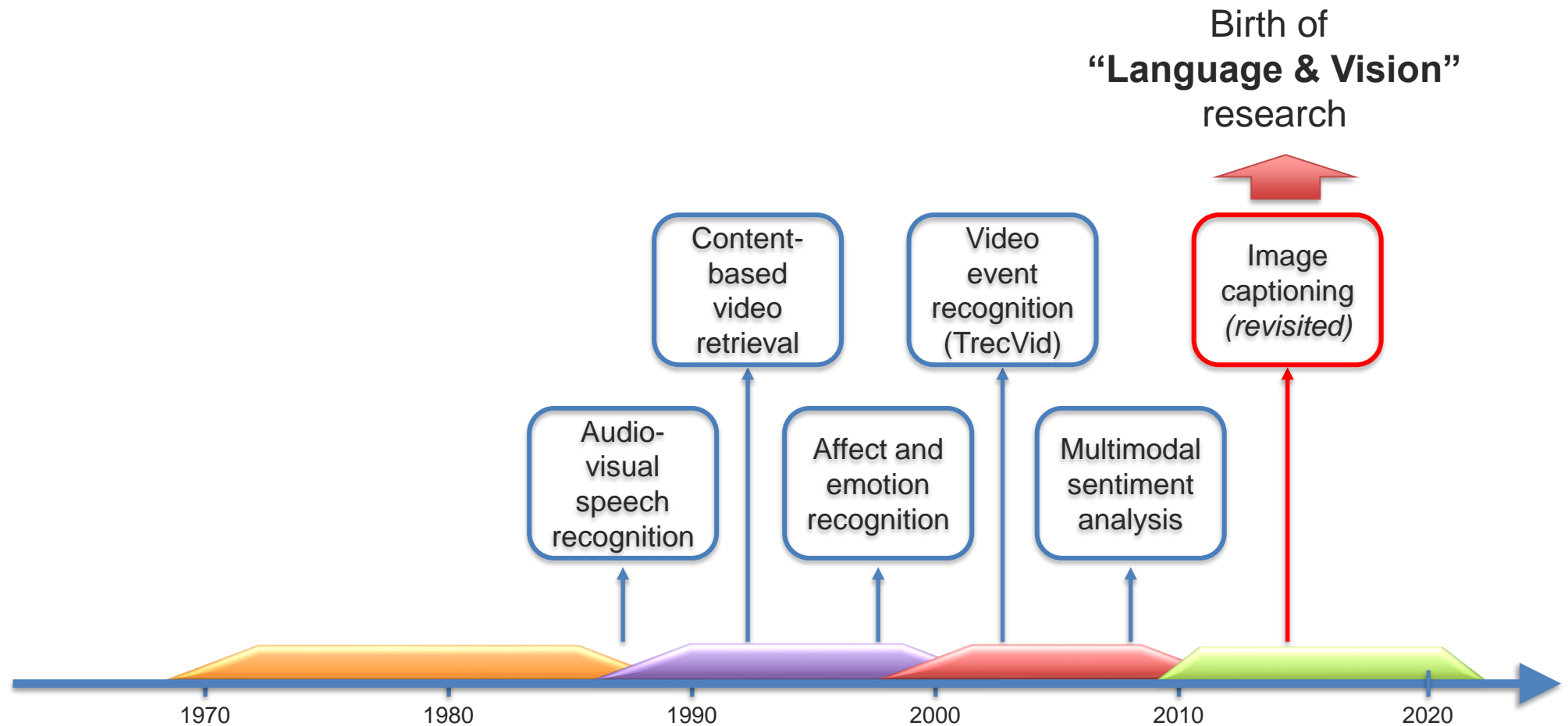
- Multimodal deep learning [ICML 2011]
- Multimodal Learning with Deep Boltzmann Machines [NIPS 2012]
- Visual attention: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [ICML 2015]

Key enablers for multimodal research:

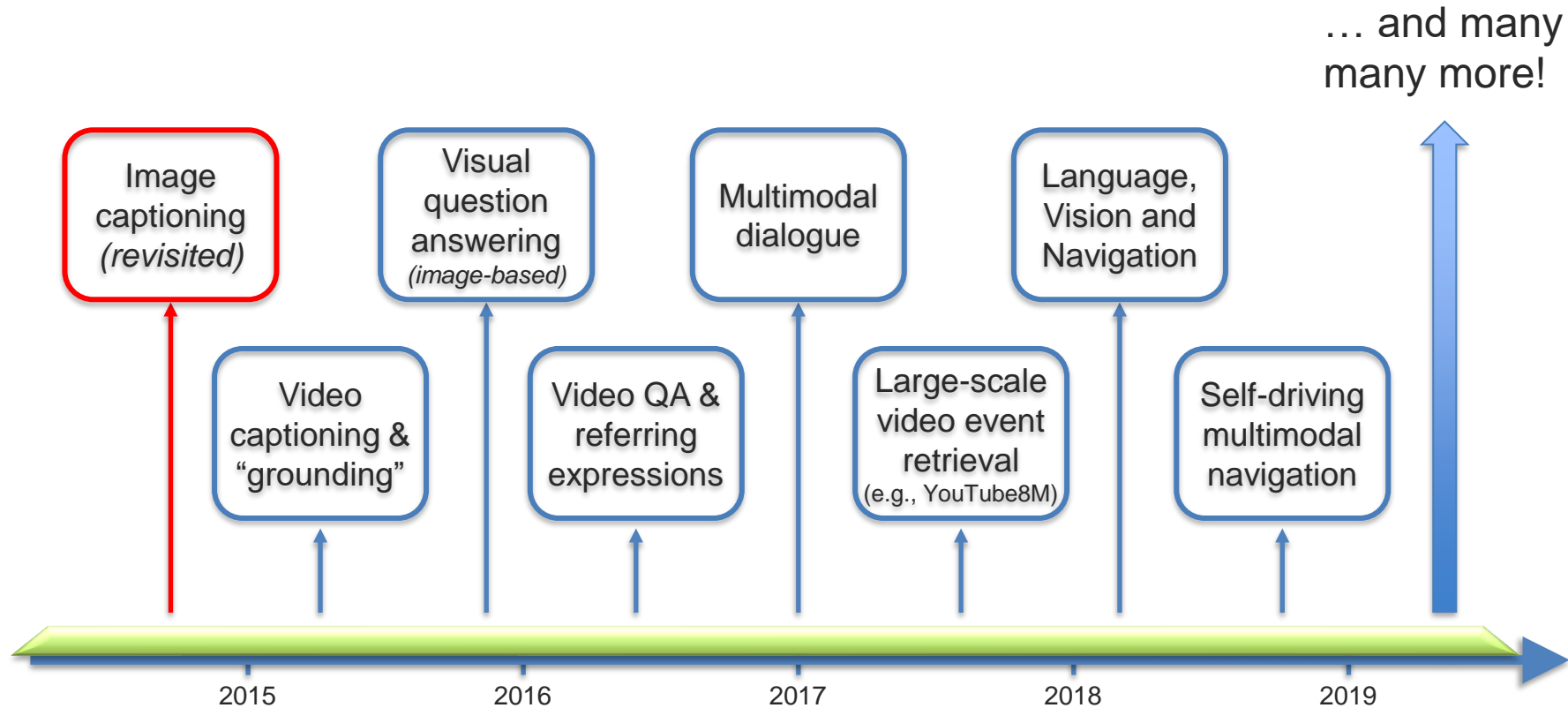
- New large-scale multimodal datasets
- Faster computer and GPUS
- High-level visual features
- “Dimensional” linguistic features



Multimodal Research Tasks



Multimodal Research Tasks



Real world tasks tackled by Multimodal ML



A. Affect recognition

- Emotion
- Personalities
- Sentiment

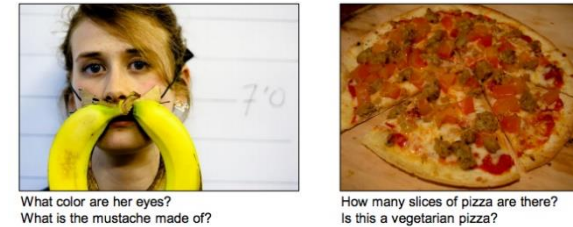


B. Media description

- Image and video captioning

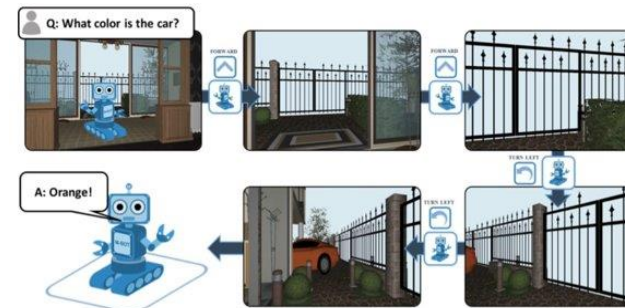
C. Multimodal QA

- Image and video QA
- Visual reasoning



D. Multimodal Navigation

- Language guided navigation
- Autonomous driving



Real world tasks tackled by Multimodal ML

E. Multimodal Dialog

- Grounded dialog

F. Event recognition

- Action recognition
- Segmentation

G. Multimedia information retrieval

- Content based/Cross-media



Multimodal Datasets

Our Latest List of Multimodal Datasets

1	Good reference dataset, but maybe not as suited for a course project
2	Maybe a good idea, but look also at alternatives.
3	Usually more recent datasets, well-suited for the project

A. Affect Recognition

AFEW	A1	1
AVEC	A2	1
IEMOCAP	A3	1
POM	A4	3
MOSI	A5	3
CMU-MOSEI	A6	3
TUMBLR	A7	2
AMHUSE	A8	1
VGD	A9	3
Social-IQ	A10	3
MELD	A11	3
MUSTARD	A12	3
DEAP	A13	3
MAHNOB	A14	3
Continuous LIRIS-ACCEDE	A15	2
DECAF	A16	2
ASCERTAIN	A17	2
AMIGOS	A18	2
EMOTIC	A19	3
M2H2	A20	3
UR-Funny	A21	3
CH-SIMS	A22	3
MuSe-CaR	A23	2
MEmoR	A24	2

B. Media Description

MSCOCO	B1	1
MPII	B2	2
MONTREAL	B3	2
LSMDC	B4	2
CHARADES	B5	3
REFEXP	B6	3
GUESSWHAT	B7	3
FLICKR30K	B8	1
CSI	B9	1
MIT-MIT	B10	3
MVSQ	B11	2
NeuralWalker	B12	2
Visual Relation	B13	3
Visual Genome	B14	3
Pinterest	B15	2
Movie Graph	B16	3
nocaps	B17	3
CrossTask	B18	2
Refer360	B19	3
Towers of Babel (WikiScenes)	B20	3
N24News	B21	2
Localized Narratives	B22	3

Our Latest List of Multimodal Datasets

1	Good reference dataset, but maybe not as suited for a course project
2	It may be a good idea, but also look at alternatives.
3	Usually more recent datasets, well-suited for the project

C. Multimodal QA

VQA	C1	1
DAQUAR	C2	1
COCO-QA	C3	2
MADLIBS	C4	2
TEXTBOOK	C5	3
VISUAL7W	C6	3
TVQA	C7	3
VCR	C8	3
Cornell NLVR	C9	3
Cornell NLVR2	C10	3
CLEVR	C11	3
EQA	C12	3
TextVQA	C13	3
GQA	C14	3
CompGuessWhat	C15	3
DVD	C16	2
AGQA	C17	3
VizWiz	C18	3
SUTD-TrafficQA	C19	3
WebQA	C20	3

D. Multimodal Navigation

Room-2-Room (R2R)	D1	1
RERERE	D2	2
VNLA	D3	3
nuScenese	D4	3
Waymo	D5	3
CARLA	D6	1
Argoverse	D7	3
ALFRED	D8	2
TEACH	D9	2
Room-across-room (RxR)	D10	3
Winoground	D11	3

Our Latest List of Multimodal Datasets

1	Good reference dataset, but maybe not as suited for a course project	
2	It may be a good idea, but also look at alternatives.	
3	Usually more recent datasets, well-suited for the project	

E. Multimodal Dialog

VISDIAL	E1	3
Talk the Walk	E2	3
Vision-and-Dialog Navigation	E3	3
CLEVR-Dialog	E4	2
Fashion Retrieval	E5	2
MMD	E6	1

F. Event Understanding

WHATS-COOKING	F1	1
TACOS	F2	2
TACOS-MULTI	F3	2
YOU-COOK	F4	1
MED	F5	1
TITLE-VIDEO-SUMM	F6	2
MEDIA-EVAL	F7	3
CRISSMMD	F8	3
EPIC-KITCHENS	F9	2
Fakedit	F10	2

G. Cross-media Retrieval

IKEA	G1	3
MIRFLICKR	G2	3
NUS-WIDE	G3	1
YAHOO-FLICKR	G4	1
YOUTUBE-8M	G5	2
YOUTUBE-BOUNDING	G6	2
YOUTUBE-OPEN	G7	2
VIST	G8	3
Recipe1M+	G9	3
VATEX	G10	3

... and please let us know (via Piazza) when you find more!

A Curated List of Multimodal Datasets

- MOSEI: Sentiment and Emotion (A6)
- Social-IQ: Modeling Social Interaction (A10)
- MELD: multi-party dialogue and emotions (A11, E)
- TVQA: Video Understanding (C7)
- NLVR2: Natural Language Grounding & Reasoning (C10)
- WebQA: Multi-hop visual and text reasoning (C20)
- Room-Across-Room: Navigation (D10)
- Winoground: Compositionality (D11)
- IKEA: multimodal retrieval (G1)

But please explore other datasets as well!!

Affect recognition dataset 2 (A2)

- Three AVEC challenge datasets 2011/2012, 2013/2014, 2015, 2016, 2017, 2018
- Audio-Visual emotion recognition
- Labeled for dimensional emotion (per frame)
- 2011/2012 has transcripts
- 2013/2014/2016 also includes depression labels per subject
- 2013/2014 reading specific text in a subset of videos
- 2015/2016 includes physiological data
- 2017/2018 includes depression/bipolar



AVEC 2011/2012



AVEC 2013/2014



AVEC 2015/2016

Multimodal Sentiment Analysis (A6)

- Multimodal sentiment and emotion recognition
- [CMU-MOSEI](#) : 23,453 annotated video segments from 1,000 distinct speakers and 250 topics

*And he I don't think he got mad when hah
I don't know maybe.*

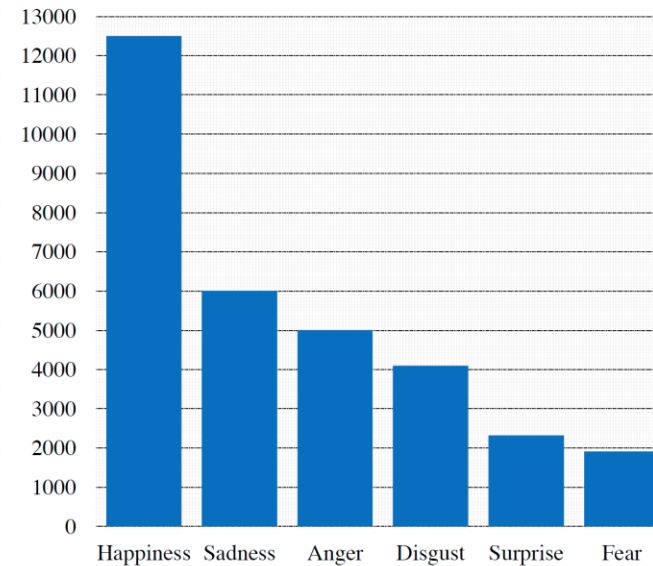
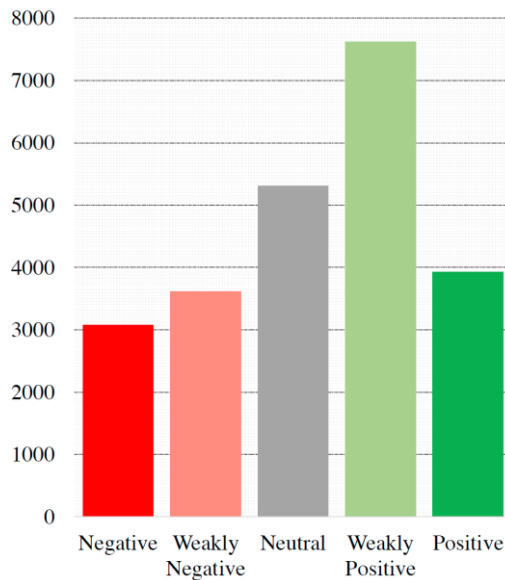


(frustrated voice)

All I can say is he's a pretty average guy.



(disappointed voice)



Media description dataset 1 – MS COCO (B1)

- Microsoft Common Objects in COntext ([MS COCO](#))
- 120000 images
- Each image is accompanied with five free form sentences describing it (at least 8 words)
- Sentences collected using crowdsourcing (Mechanical Turk)
- Also contains object detections, boundaries and keypoints



The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.

Visual Questions & Answers – VQA (C1)

- Task - Given an image and a question, answer the question (<http://www.visualqa.org/>)



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

Social Interaction Q&A Dataset (A10)

- [Social-IQ](#): 1.2k videos, 7.5k questions, 50k answers
- Questions and answers centered around social behaviors

00:29 → 00:37 00:37 → 00:40 00:40 → 00:42

(trying to speak) Steven went, got the keys and we gonna have them back. That easy. (serious face)

I couldn't ... (Interrupts) But this was Friday Matt! This was Friday. (serious face)

(silenced) You said you were going to do it and you are not doing it!

Q1: How is the discussion between the woman and the man in the white shirt ? <intermediate>
A1. The woman is blaming the man in the white shirt who seems to be in the fault. <easy>
A2. She is blaming her in a tense voice and not letting him defend himself. <advanced>
A3. They are having a romantic conversation. <easy>
A4. An active argument that both are blaming each other. <advanced>

Q2: How is the man who is not being blamed responding to the situation? <advanced>
A1. He thinks the other man is slacking even if he is not saying it. <advanced>
A2. He is showing support for the woman by taking her side. <intermediate>
A3. He thinks he is better than both of the people arguing. <easy>
A4. He doesn't want to pick a side. <advanced>

Q3: Why is the woman seem so overwhelmed? <advanced>
A1. Because a small problem became a huge problem. <intermediate>
A2. She has too much on her plate, and this new problem overwhelms her. <advanced>
A3. The woman is upset because the men are insulting her. <easy>
A4. Because both of them men seem to be ignoring her. <intermediate>

Multimodal QA (C7)

- TVQA
 - Video QA dataset based on 6 popular TV shows
 - 152.5K QA pairs from 21.8K clips
 - Compositional questions

00:00:755 --> 00:02.655
(Chandler:) Go to your room!
00:06.961 --> 00:08.622
(Janice:) I gotta go, I gotta go.

00:08.829 --> 00:10.057
(Janice:) Not without a kiss.
00:10.264 --> 00:12.391
(Chandler:) Maybe I won't kiss you so you'll stay.

00:12.600 --> 00:14.761
(Joey:) Kiss her. Kiss her!
00:16.771 --> 00:19.137
(Janice:) I'll see you later, sweetie. Bye, Joey.

...

00:39.327 --> 00:40.760
(Chandler:) She makes me happy.
00:41.596 --> 00:44.087
(Joey:) Okay. All right.

00:00 00:06 00:10 00:17 00:39 00:45 01:04

What is Janice holding on to **after Chandler sends Joey to his room?**

A Chandler's tie
B Chandler's hands
C Her Breakfast
D Her coat
E Chandler's coffee cup.

Why does Joey want Chandler to kiss Janice **when they are in the kitchen?**

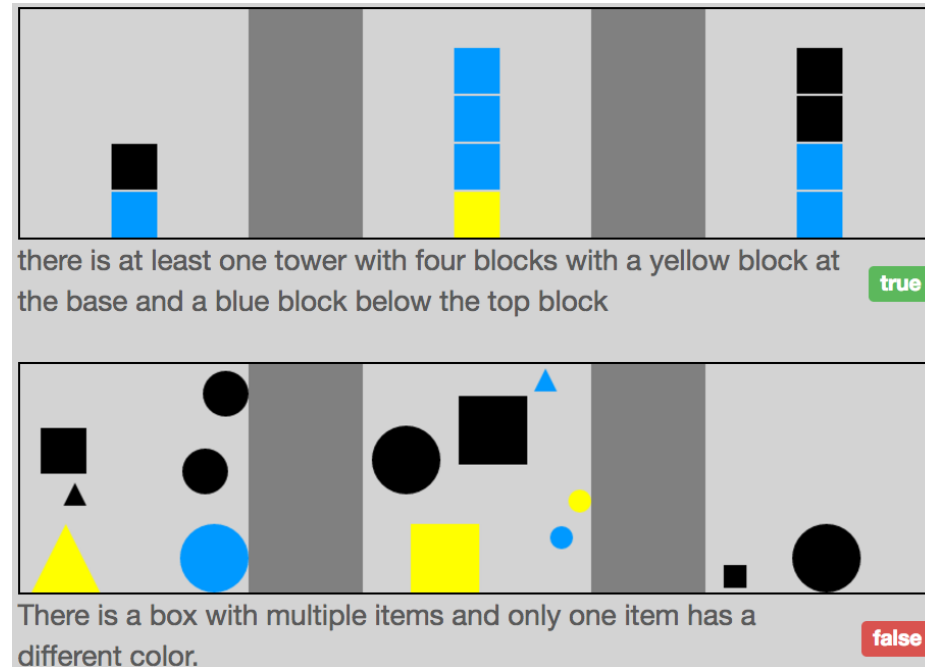
A Because Joey is glad that Chandler is happy
B Because Joey likes to watch people kiss
C **Because then she will leave**
D Because Joey thinks Janice is hot
E Because then Chandler will move away from the toast.

What is on the couch behind Joey **when he is at the counter?**

A A chick
B **A soccer ball**
C A duck
D A pillow
E Janice's coat

Multimodal QA – Visual Reasoning (C9)

- [Cornell NLVR](#)
 - 92,244 pairs of natural language statements grounded in synthetic images
 - Determine whether a sentence is true or false about an image



Multimodal QA – Visual Reasoning (C10)

- [Cornell NLVR2](#)
 - Same as NLVR but with >100k real images



The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.



One image shows exactly two brown acorns in back-to-back caps on green foliage.

WebQA (C20)

- <https://webqna.github.io/>
- Given a question Q, and a list of sources S = {s1, s2, ...}, a system must a) identify the sources from which to derive the answer, and b) generate an answer as a complete sentence.

Q: At which festival can you see a castle in the background: Oktoberfest in Domplatz Austria or Tanabata festival in Hiratsuka, Japan?

J24 029 Dom, Oktoberfest

The festival is a "Syonan Hiratsuka Tanabata Matsuri".

Tanabata festival in Hiratsuka

Calella - Catalonia, Spain - 11 Aug. 2009

For the Oktoberfest, Löwenbräu brews a special Märzen beer called Oktoberfestbier or Wiesenbier ("meadow beer," referring to the Bavarian name of the festival site, the "Wiesn").

In the summer, the Sendai Tanabata Festival, the largest Tanabata festival in Japan, is held. In winter, the trees are decorated with thousands of lights for the Pageant of Starlight, lasting through most of December.

Maskkrone Four mugs of beer at Oktoberfest 2008.

Fussa Tanabata Festival-Tokyo

Ghost train on the Munich Oktoberfest.

A: You can see a castle in the background at Oktoberfest in Domplatz, Austria

Navigating in a Virtual House (D1)

Visually-grounded natural language navigation in real buildings

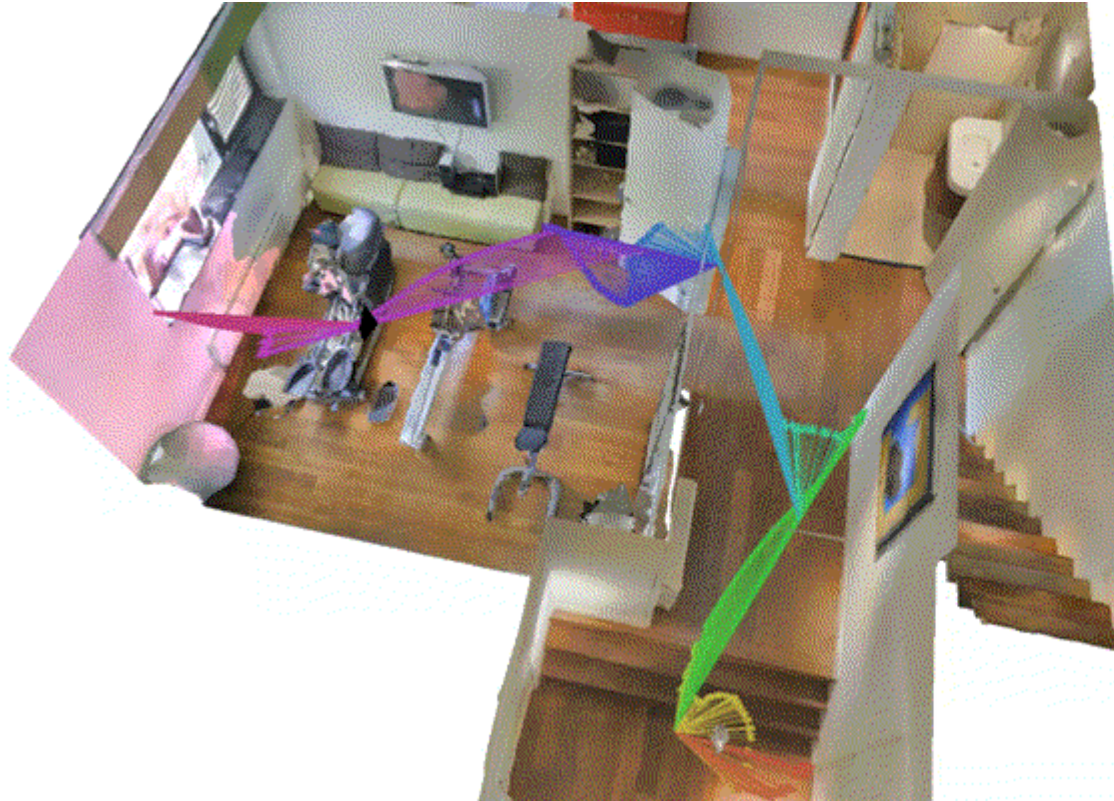
- [Room-2-Room](#): 21,567 open vocabulary, crowd-sourced navigation instructions



Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

Room-Across-Room (D10)

- [Github](#)
- Similar to Room-to-Room (D1) except larger, multilingual, and with longer paths



Now you are standing in front of a closed door, turn to your left, you can see two wooden steps, climb the steps and walk forward by crossing a wall paint which is to your right side, you can see open door enter into it. This is a gym room, move forward, walk till the end of the room, you can see a grey colored ball to the corner of the room, stand there, that's your end point.

Winoground (D11)

- [Github](#)
- Same words, different order, different images. Intended to test the compositionality of vision-language models



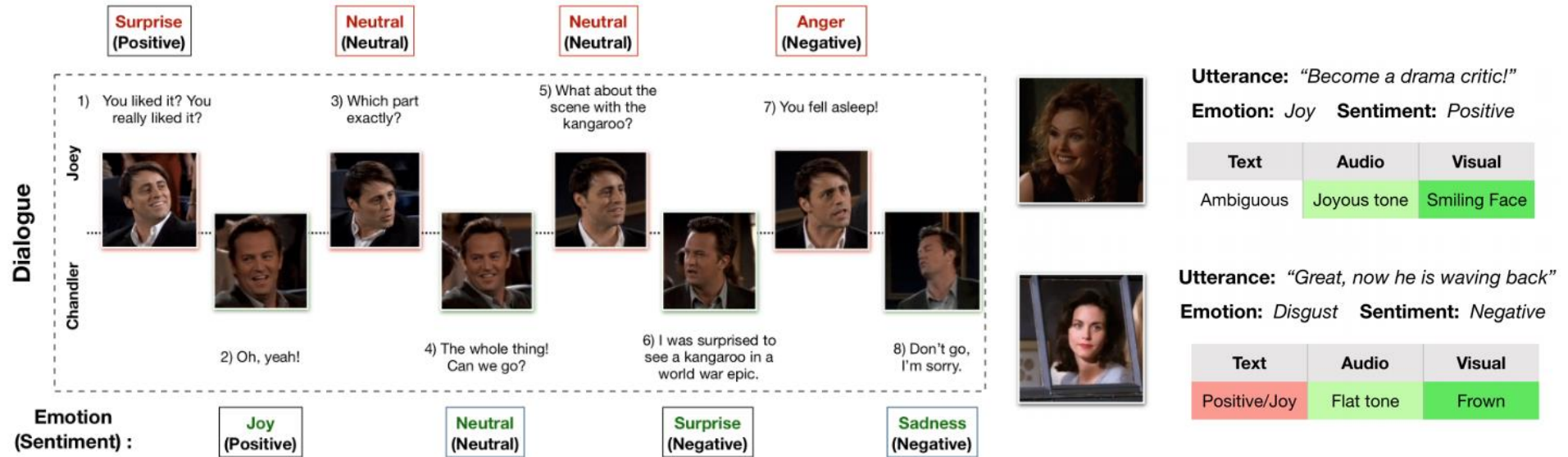
(a) some plants surrounding a lightbulb



(b) a lightbulb surrounding some plants

Multi-Party Emotion Recognition (A11, E)

- [MELD](#): Multi-party dataset for emotion recognition in conversations



EPIC-Kitchens (F9)

- [Dataset](#)
- Large-scale dataset in first-person (egocentric) vision; multi-faceted, audio-visual, non-scripted recordings in native environments - i.e. the wearers' homes



Multimodal Retrieval: IKEA Interior Design Dataset (G1)

- [Interior Design Dataset](#) – Retrieve desired product using room photos and text queries.
- 298 room photos, 2193 product images/descriptions.

Room images:



Object images: Description:



You sit comfortably thanks to the armrests.



There's a natural and living feeling of wood, as knots and other marks remain on the surface.



This lamp gives a pleasant light for dining and spreads a good directed light across your dining or bar table.

Some Advice About Multimodal Datasets

- If you are used to deal with text or speech
 - Space will become an issue working with image/video data
 - Some datasets are in 100s of GB (compressed)
- Memory for processing it will become an issue as well
 - Won't be able to store it all in memory
- Time to extract features and train algorithms will also become an issue
- Plan accordingly!
 - Sometimes tricky to experiment on a laptop (might need to do it on a subset of data)

Available Tools

- Use available tools in your research groups
 - Or pair up with someone that has access to them
- Find some GPUs!
- We will be getting AWS credit for some extra computational power
- Google Cloud Platform credit as well



Upcoming Course Assignments

Project preferences (deadline Tuesday 9/6 at 8pm ET)

- Let us know about your project preferences, including datasets, research topics and potential teammates
 - See instructions on Piazza
- We will reserve a moment for discussions on Thursday 9/8 to help you with finding project teammates

Reading Assignment (Summaries due Friday 9/9 at 8pm ET)

- We created the study groups in Piazza.
 - End of the discussion period: Monday 9/12 at 8pm ET

Lecture Highlights (for both lectures next week)

- Starting next week, you need to post your lecture highlights following each course lecture. See Piazza for detailed instructions.

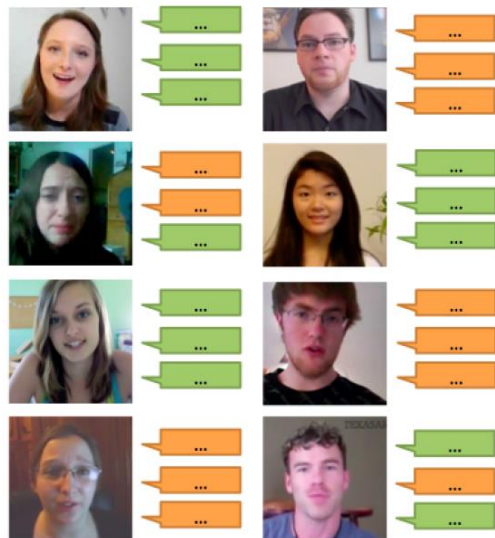
Examples of Previous Projects

Project Example: Select-Additive Learning

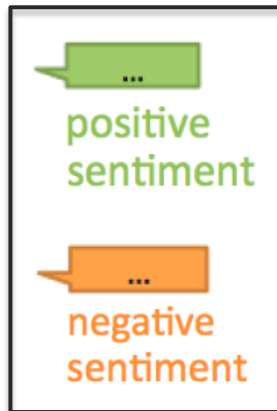
Research task: Multimodal sentiment analysis

Datasets: MOSI, YouTube, MOUD

Main idea: Reducing the effect of *confounding factors* when limited dataset size



Legend



What rules can you infer from this data?

- ✓ Smile -> positive sentiment
- ✓ Frown -> negative sentiment
- ✓ nod -> positive sentiment
- ✗ Wearing glasses -> negative sentiment

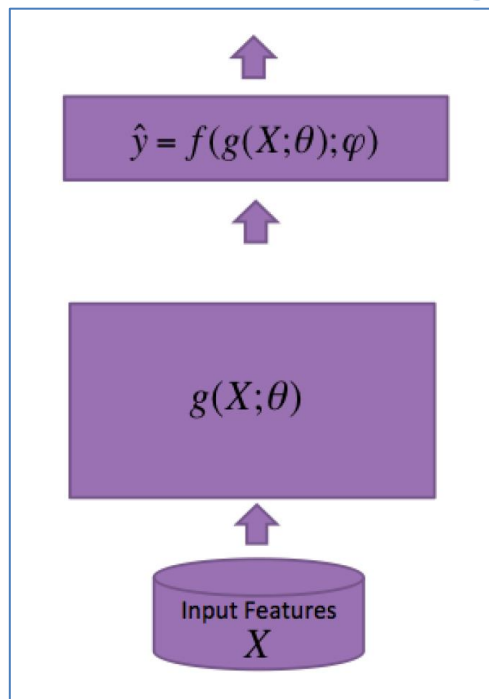
Confounding factor!

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency and Eric P. Xing, Select-additive Learning: Improving Generalization In Multimodal Sentiment Analysis, ICME 2017, <https://arxiv.org/abs/1609.05244>

Project Example: Select-Additive Learning

Solution: Learning representations that reduce the effect of user identity

“Conventional”
representation learning



Select-Additive Learning



Hypothesis: the representation is a mixture from the person-independent factor $g(X)$ and the person-dependent factor $h(Z)$.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency and Eric P. Xing, Select-additive Learning: Improving Generalization In Multimodal Sentiment Analysis, ICME 2017, <https://arxiv.org/abs/1609.05244>

Project Example: Word-Level Gated Fusion

Research task: Multimodal sentiment analysis

Datasets: MOSI, YouTube, MOUD

Main idea: Estimating importance of each modality at the word-level in a video.



Visual Gate:

Reject

Pass

Reject



Visual modality: Hands cover mouth

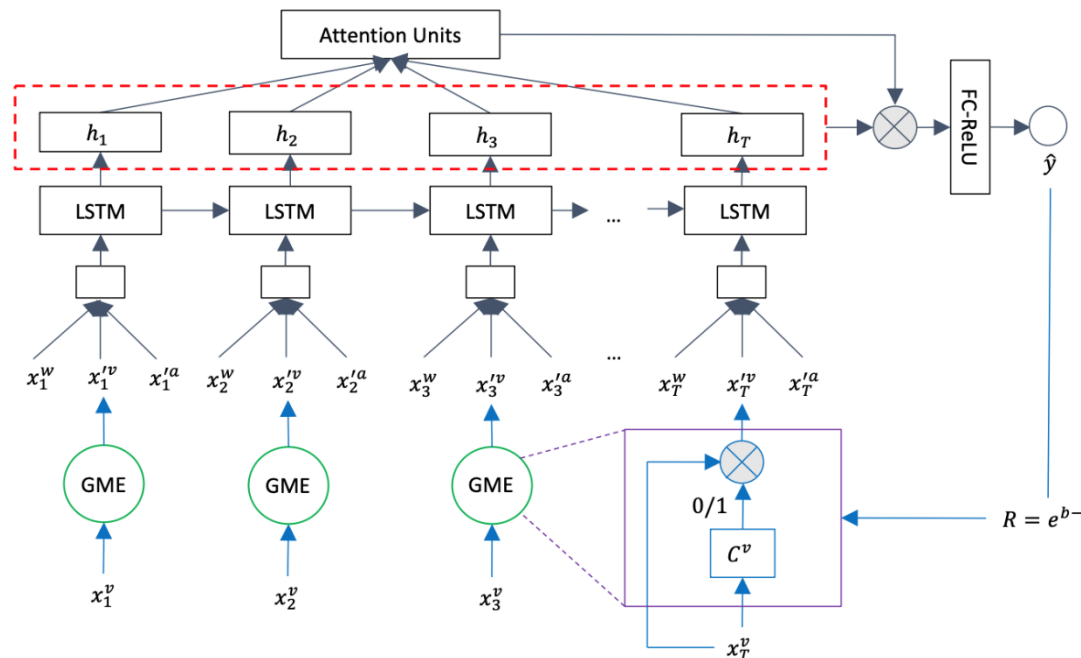
How can we build an interpretable model that estimates modality and temporal importance, and learns to attend to important information?

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, Louis-Philippe Morency, Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning, ICMI 2017, <https://arxiv.org/abs/1802.00924>

Project Example: Word-Level Gated Fusion

Solution:

- Word-level alignment
- Temporal attention over words
- Gated attention over modalities



Hypothesis: attention weights represent contribution of each modality at each time step

Modality gates that determine importance and contribution of each modality – trained with reinforcement learning

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, Louis-Philippe Morency, Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning, ICMI 2017, <https://arxiv.org/abs/1802.00924>

Project Example: Instruction Following

Research task: Task-Oriented Language Grounding in an Environment

Datasets: ViZDoom, based on the Doom video game

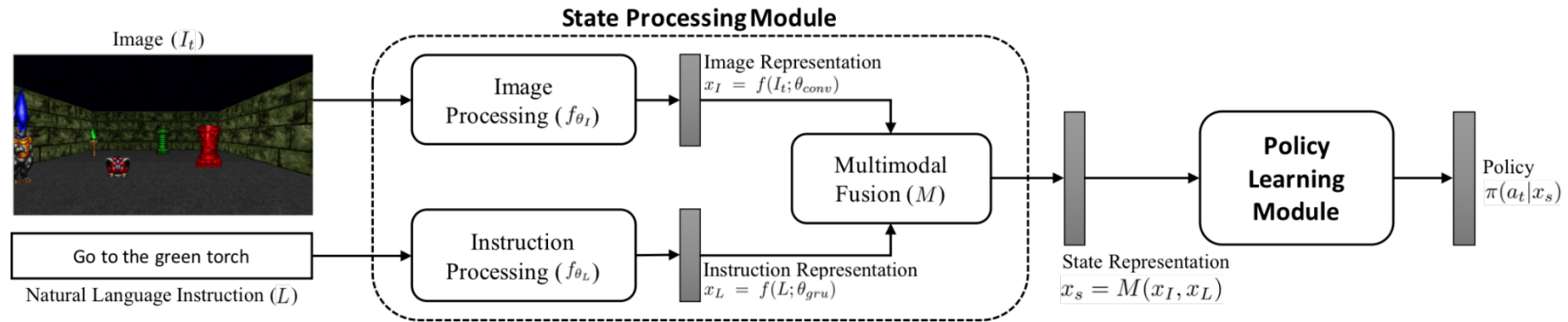
Main idea: Build a model that comprehends natural language instructions, grounds the entities and relations to the environment, and execute the instruction.



Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, Ruslan Salakhutdinov, Gated-Attention Architectures for Task-Oriented Language Grounding. AAAI 2018 <https://arxiv.org/abs/1706.07230>

Project Example: Instruction Following

Solution: Gated attention architecture to attend to instruction and states



Hypothesis: Gated attention learns to ground and compose attributes in natural language with the image features. e.g. learning grounded representations for 'green' and 'torch'.

Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, Ruslan Salakhutdinov, Gated-Attention Architectures for Task-Oriented Language Grounding. AAAI 2018 <https://arxiv.org/abs/1706.07230>

Project Example: Adversarial Attacks on VQA models

Research task: Adversarial Attacks on VQA models

Datasets: VQA

Main idea: Test the robustness of VQA models to adversarial attacks on the image.



“panda”

57.7% confidence

+ .007 ×



noise

=



“gibbon”

99.3% confidence

Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, Louis-Philippe Morency, Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models. NeurIPS ViGIL workshop 2018. <https://nips2018vigil.github.io/static/papers/accepted/33.pdf>

Project Example: Adversarial Attacks on VQA models

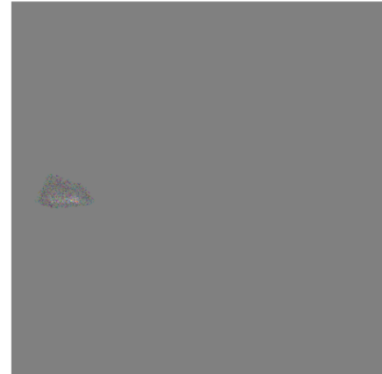
Research task: Adversarial Attacks on VQA models

Datasets: VQA

Main idea: Test the robustness of VQA models to adversarial attacks on the image.



+



Q: what kind of flowers are in the vase?



VQA model



A: **Roses** to **Sunflower**

How can we design a targeted attack on images in VQA models, which will help in assessing robustness of existing models?

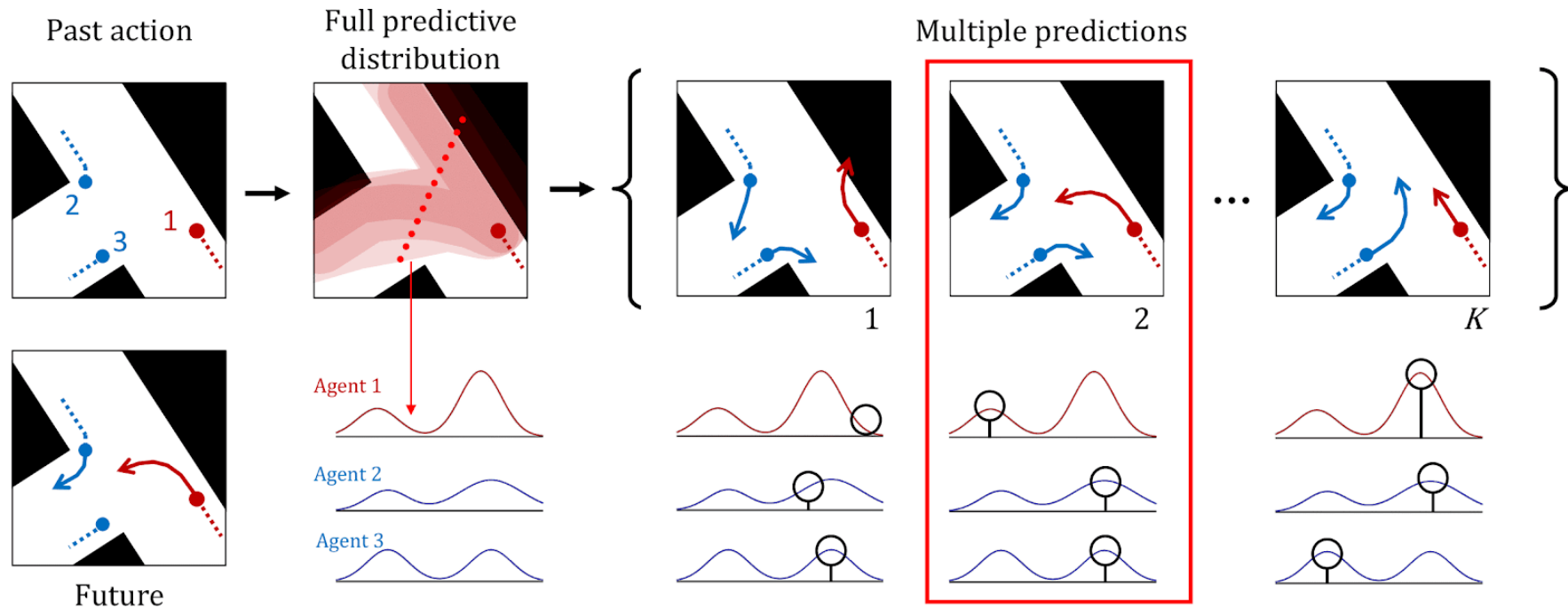
Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, Louis-Philippe Morency, Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models. NeurIPS ViGIL workshop 2018. <https://nips2018vigil.github.io/static/papers/accepted/33.pdf>

Project Example: Multiagent Trajectory Forecasting

Research task: Multiagent trajectory forecasting for autonomous driving

Datasets: Argoverse and Nuscenes autonomous driving datasets

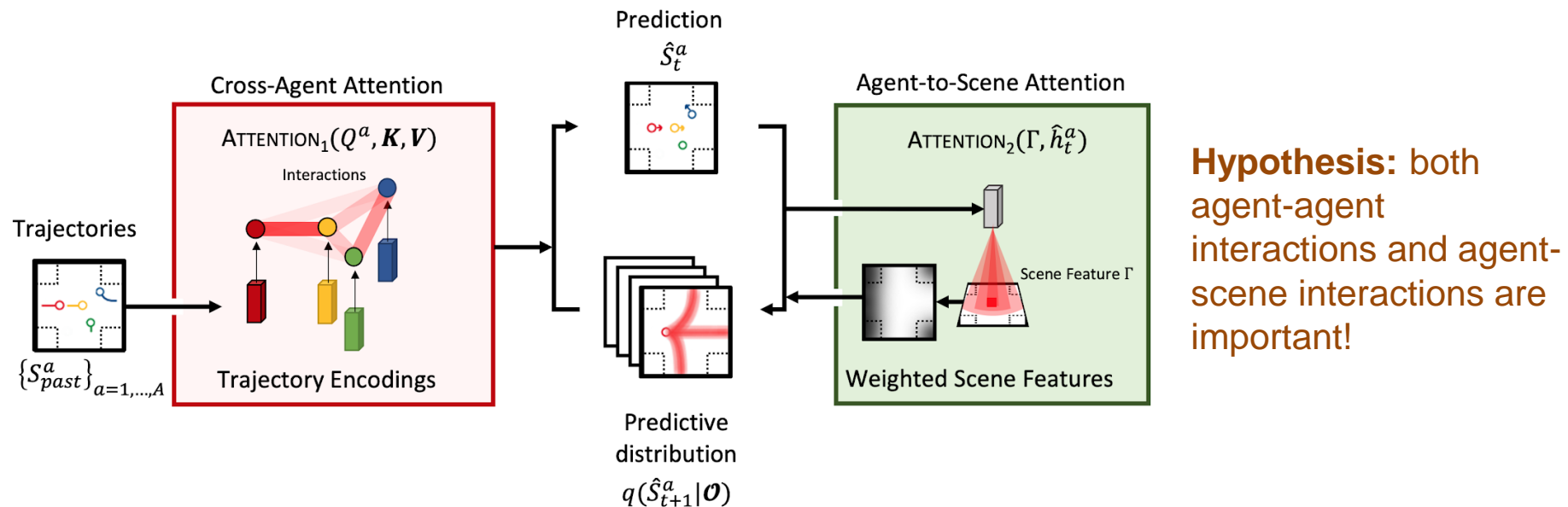
Main idea: Build a model that understands the environment and multiagent trajectories and predicts a set of multimodal future trajectories for each agent.



Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R. Jadhav, Paul Pu Liang, Louis-Philippe Morency, Diverse and Admissible Trajectory Forecasting through Multimodal Context Understanding. ECCV 2020 <https://arxiv.org/abs/1706.07230>

Project Example: Multiagent Trajectory Forecasting

Solution: Modeling the environment and multiple agents to learn a distribution of future trajectories for each agent.

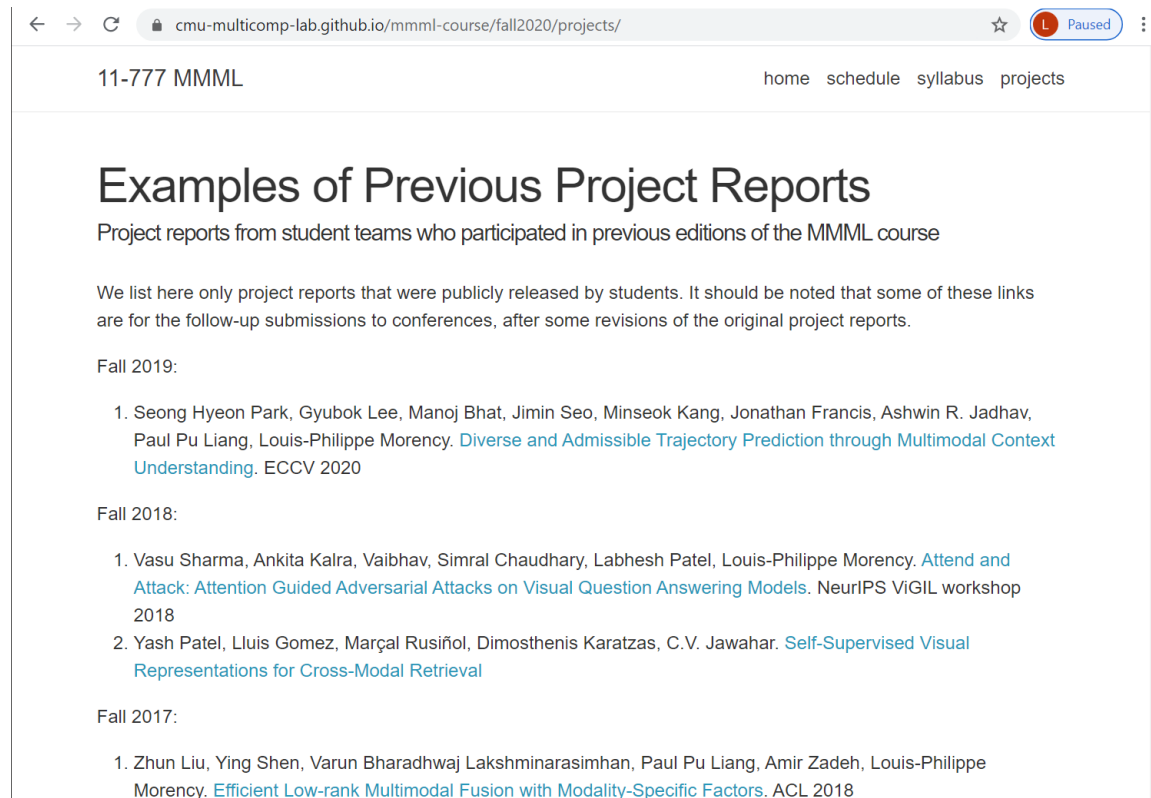


Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R. Jadhav, Paul Pu Liang, Louis-Philippe Morency, Diverse and Admissible Trajectory Forecasting through Multimodal Context Understanding. ECCV 2020
<https://arxiv.org/abs/1706.07230>

More Project Examples

See the Fall 2020 course website:

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



The screenshot shows a web browser window displaying the course website. The address bar shows the URL <https://cmu-multicomp-lab.github.io/mmml-course/fall2020/projects/>. The page title is "11-777 MMML" and the navigation menu includes "home", "schedule", "syllabus", and "projects". The main heading is "Examples of Previous Project Reports". Below the heading, there is a paragraph explaining that the page lists project reports from student teams who participated in previous editions of the MMML course, noting that some links are for follow-up submissions to conferences. The page is organized by year, with sections for Fall 2019, Fall 2018, and Fall 2017. Each section contains a list of project reports with their authors and titles.

11-777 MMML home schedule syllabus projects

Examples of Previous Project Reports

Project reports from student teams who participated in previous editions of the MMML course

We list here only project reports that were publicly released by students. It should be noted that some of these links are for the follow-up submissions to conferences, after some revisions of the original project reports.

Fall 2019:

1. Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R. Jadhav, Paul Pu Liang, Louis-Philippe Morency. [Diverse and Admissible Trajectory Prediction through Multimodal Context Understanding](#). ECCV 2020

Fall 2018:

1. Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, Louis-Philippe Morency. [Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models](#). NeurIPS ViGIL workshop 2018
2. Yash Patel, Lluís Gomez, Marçal Rusiñol, Dimosthenis Karatzas, C.V. Jawahar. [Self-Supervised Visual Representations for Cross-Modal Retrieval](#)

Fall 2017:

1. Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency. [Efficient Low-rank Multimodal Fusion with Modality-Specific Factors](#). ACL 2018