Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 1.2: Multimodal Research Tasks

Paul Liang

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

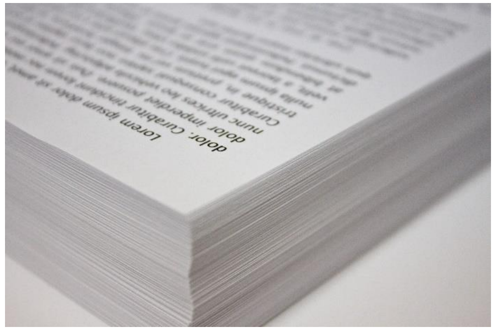
Lecture Objectives

- Course syllabus and project assignments
  - Course and assignment schedule
  - Projects and team matching
  - Grades and course structure
- Experimental design
  - Research questions and hypotheses
- A historical view on multimodal research
- 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)

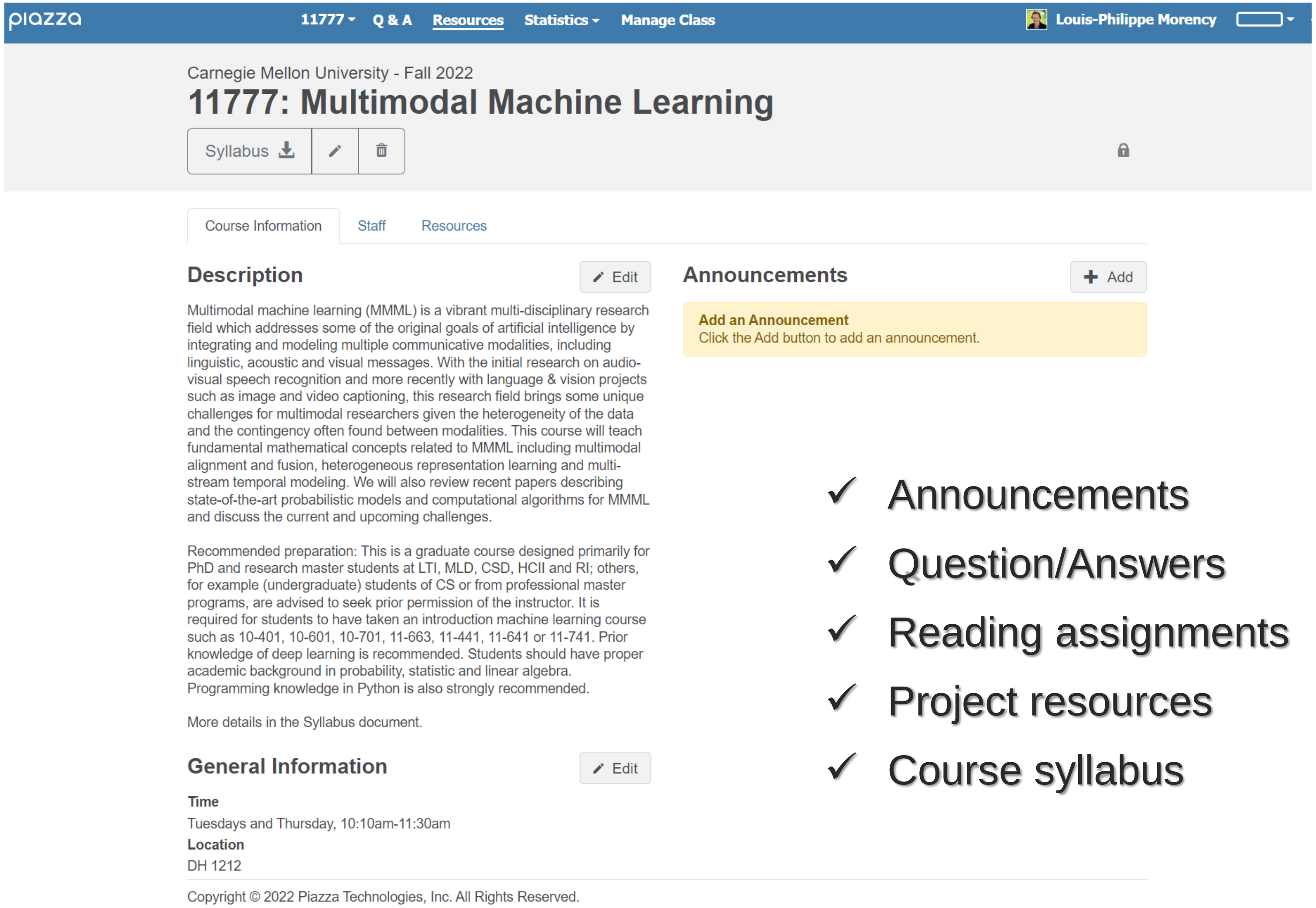


Reading assignments
(12% of your grade)

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

Course project assignments
(72% of your grade)

Piazza https://piazza.com/cmu/fall2023/11777/info



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

Website https://cmu-multicomp-lab.github.io/mmml-course/fall2023/

11-777 MMML

home schedule readings syllabus projects

MultiModal Machine Learning

11-777 • Fall 2023 • 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 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.

The course will present the fundamental mathematical concepts in machine learning and deep learning relevant to the six main challenges in multimodal machine learning: (1) representation, (2) alignment, (3) reasoning, (4) generation, (5) transference and (6) quantification. These include, but not limited to, multimodal transformers, neuro-symbolic models, multimodal tensor fusion, mutual information and multimodal graph networks. The course will also discuss many of the recent applications of MMML including multimodal affect recognition, multimodal language grounding and language-vision navigation.

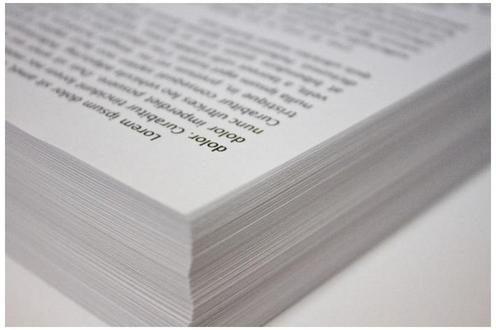
- **Time:** Tuesday and Thursday 9:30-11:00 AM
- **Content:** CMU Canvas
- **Location:** MM A14 and zoom (see links in CMU Canvas)
- **Discussion and Q&A:** Piazza
- **Assignment submissions:** Gradescope (for registered students only)
- **Online lectures:** The lectures will be recorded and made available on CMU Canvas for registered students. External link to the lectures on our Youtube channel!
- **Contact:** Students should ask all course-related questions on Piazza, where you will also find announcements.

Updated slower, mainly for non-CMU public to access

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



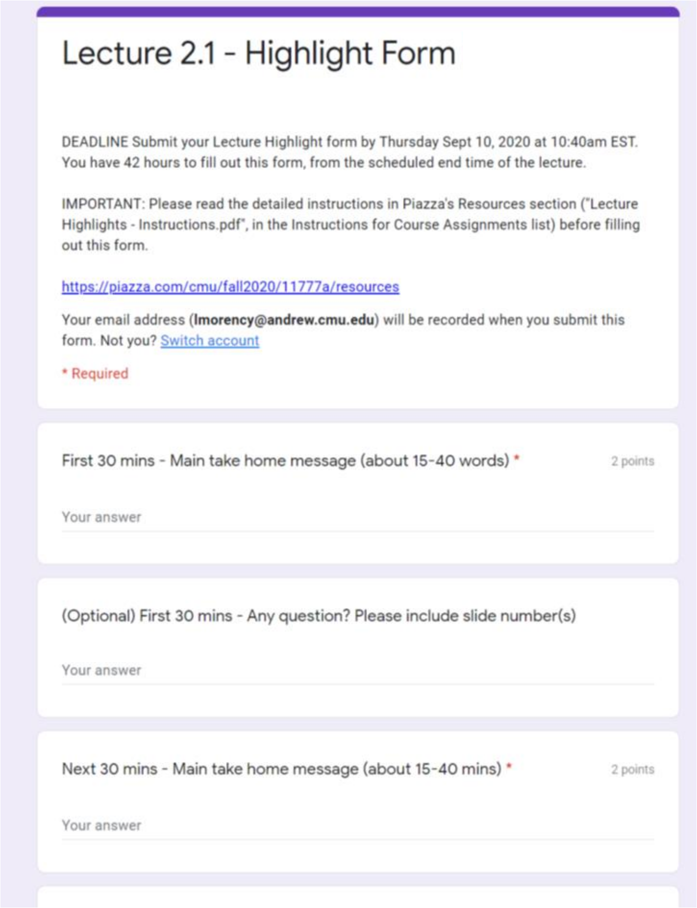
$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$
$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$
$$h_t = o_t \tanh(c_t)$$

- 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 (16%)

Starting Week 2 !!



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 (**lmorency@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)

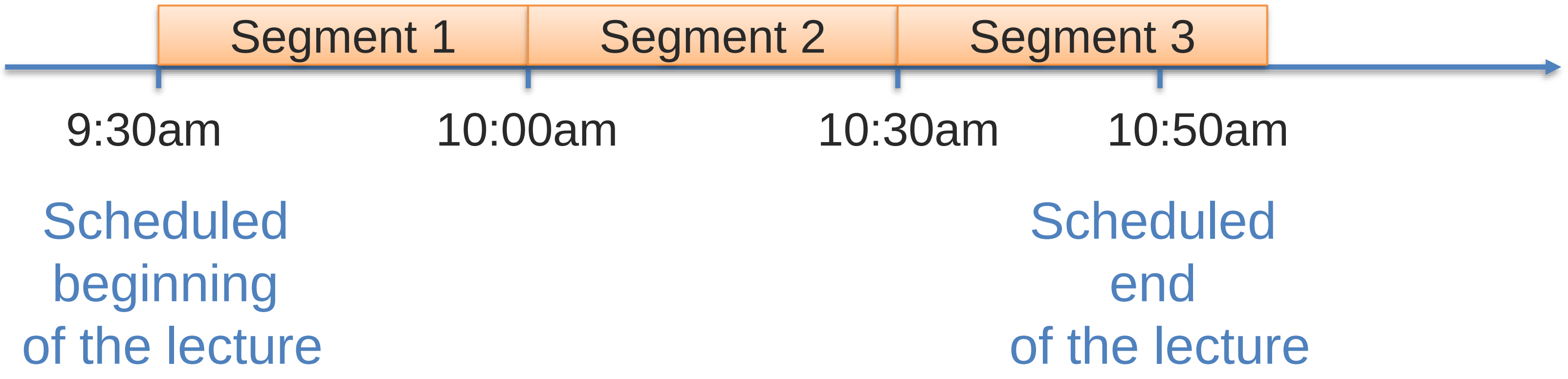
Ask questions about the lecture

- Will be answered either online or at the next lecture

Submitted same day as lecture (before 9pm)

- Students are encouraged to attend lectures in person

Lecture Highlight Form – Segments (16%)



- Segment 1 starts at 9:30am, even if the lecture starts slightly later.
- Segment 3 ends whenever the lecture ends
- Slides happening around the segment borders (+/- 5min of 10:00am and 10:30am) can be included in either neighboring segment.

First Reading Assignment – Week 2 (12%)

- Study groups: 9-10 students per group (randomly, in Piazza)
- 4 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 (12%)

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/fall2023/11777/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/fall2023/11777/resources

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 2024, ACL 2024, IJCAI 2024, ICML 2024, ICMI 2024
- We will discuss on Thursday about project ideas
- GPU resources available:
  - Amazon AWS and Google Cloud Platform

Course Project Timeline

Pre-proposal *(due Wednesday Sept. 13)*

- Define your dataset, research task and teammates

First project assignment *(due Sunday Sept. 24)*

- Study related work to your selected research topic

Second project assignment *(due Sunday Oct 8)*

- Experiment with unimodal representations

Midterm project assignment (due Sunday Oct 29)

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

Final project assignment (due Sunday Dec. 10)

- 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 8/31 (today!):** Lecture describing available multimodal datasets and research topics
- **Tuesday 9/5:** Let us know your dataset preferences for the course project
- **Thursday 9/7:** During the later part of the lecture, we will have an interactive period to help with team formation. More details to come
- **Wednesday 9/13:** Pre-proposals are due. You should have selected your teammates, dataset and task

Project Preferences – Due Tuesday 9/5

- 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)

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 1 8/29 & 8/31	**Course introduction** • Multimodal core challenges • Course syllabus	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 9/5 & 9/7 Read due: 9/9	**Unimodal representations** • Dimensions of heterogeneity • Visual representations	**Unimodal representations** • Language representations • Signals, graphs and other mod
Week 3 9/12 & 9/14 Read due: 9/16 Proj. Due: 9/13	**Multimodal representations** • Cross-modal interactions • Multimodal fusion	**Multimodal representations** • Coordinated representations • Multimodal fission
Week 4 9/19 & 9/21 *Proj. due: 9/24*	**Multimodal alignment and grounding** • Explicit alignment • Multimodal grounding	**Alignment and representations** • Self-attention transformer mo • Masking and self-supervised le
Week 5 9/26 & 9/28 Read due: 9/30	**Multimodal transformers** • Multimodal transformers • Video and graph representations	**Multimodal Reasoning** • Structured and hierarchical models • Memory models
Week 6 10/3 & 10/5 *Proj. due: 10/8*	***Project hours***	***Multimodal language grounding*** • Grounded semantics and prag

Project preferences due on Tuesday 9/5

Pre-proposals due on Wednesday 9/13

First assignment due on Sunday 9/24

Second assignment due on Sunday 10/8

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 7 10/10 & 10/12 Read due: 10/14	**Multimodal interaction** • Reinforcement learning • Discrete structure learning	**Multimodal inference** • Logical and causal inference • External knowledge
Week 8 10/17 & 10/19	***Fall Break – No lectures***	
Week 9 10/24 & 10/26 *Proj. due: 10/29*	**Multimodal generation** • Translation, summarization, creation • Generative models: VAEs	**New generative models** • GANs and diffusion model • Model evaluation and ethi
Week 10 10/31 & 11/2	***Project presentations (midterm)***	***Project presentations (midterm)***
Week 11 11/7 & 11/9 Read due: 11/12	***Democracy Day – No Class –***	**Transference** • Modality transfer and co-learning • Self-training and multitask learning
Week 12 11/14 & 11/16 Read due: 11/21	**Quantification** • Heterogeneity and interactions • Biases and fairness	**New research directions** • Recent research in multimodal ML

Midterm assignment due on Sunday 10/29

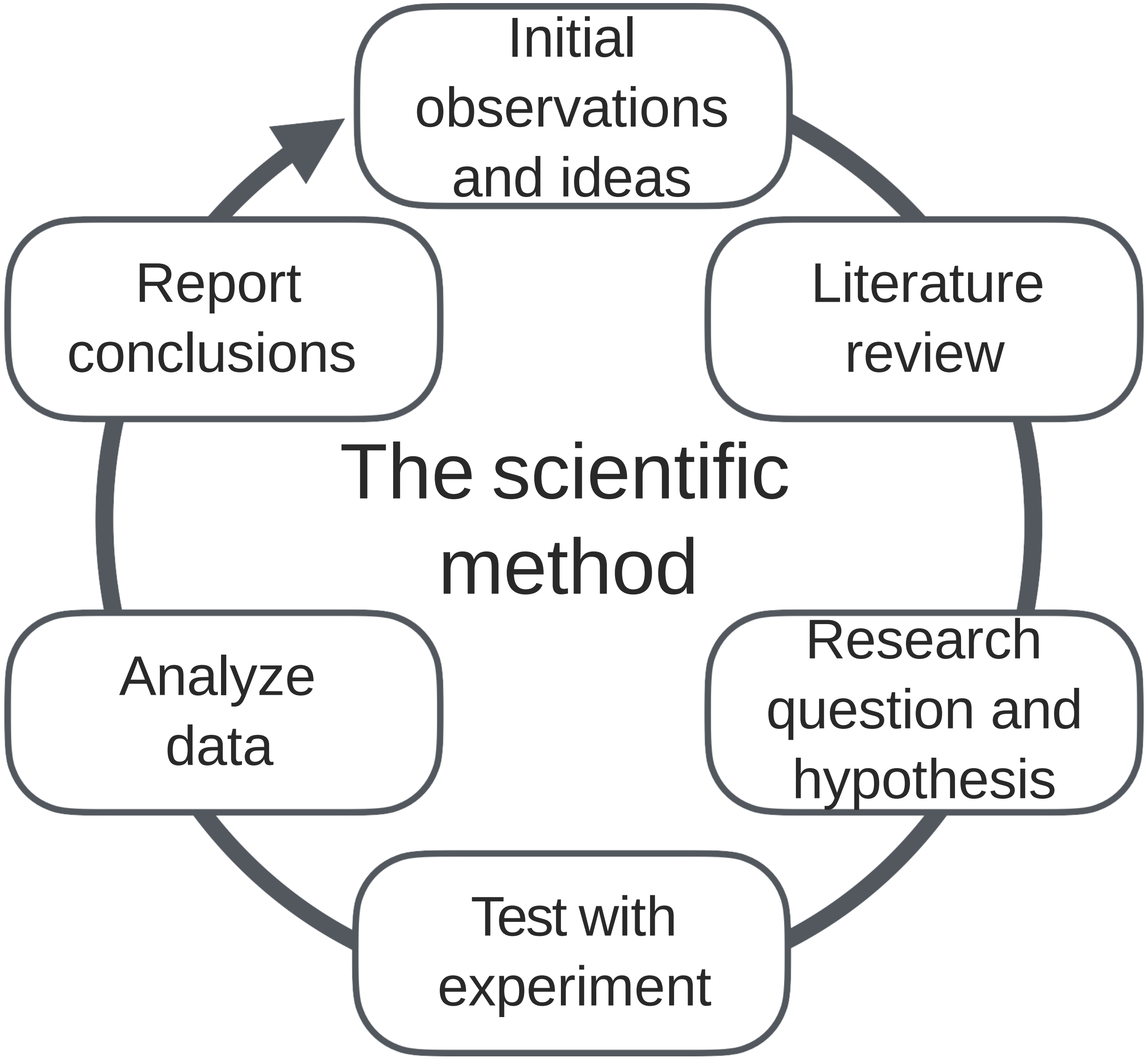
Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 13 11/21 & 11/23	***Thanksgiving Week – No Class –***	
Week 14 11/28 & 11/30	**Guest lecture**	**Guest lecture**
Week 15 12/5 & 12/7 *Proj. due: 12/10*	***Project presentations (final)***	***Project presentations (final)***

Final assignment due on Sunday 12/10

Experimental Design

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



Credit: Adapted From Wikipedia (Efbrazil)

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)

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

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

Examples of Research Ideas

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

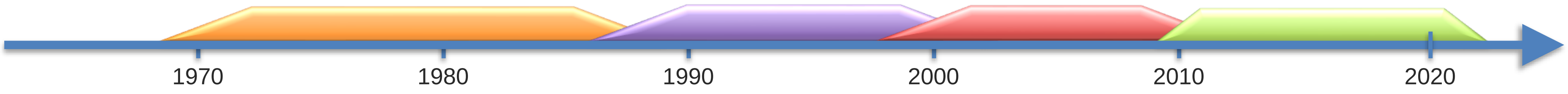
- Better understanding of the cross-modal interactions in multimodal models
- Understanding compositionality and multimodal reasoning
- Robustness to missing/noisy modalities, adversarial attacks
- Studying social biases and creating fairer models
- Interpretable and trustworthy models
- Faster and more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

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

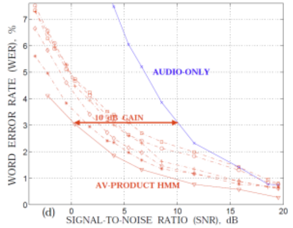
David McNeill

"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

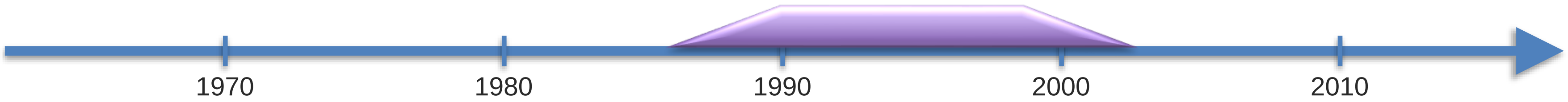
2) Multimodal interfaces

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

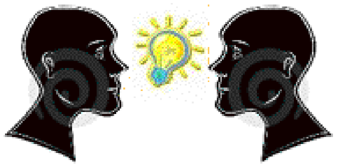
AMI Project [2001-2006, IDIAP]

- 100+ hours of meeting recordings
- Transcribed and annotated

PAL
Personalized Assistant that Learns

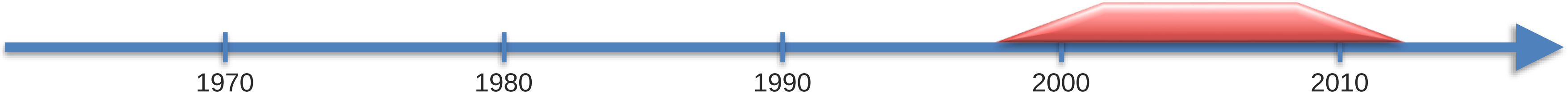
CALO Project [2003-2008, SRI]

- 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 …)

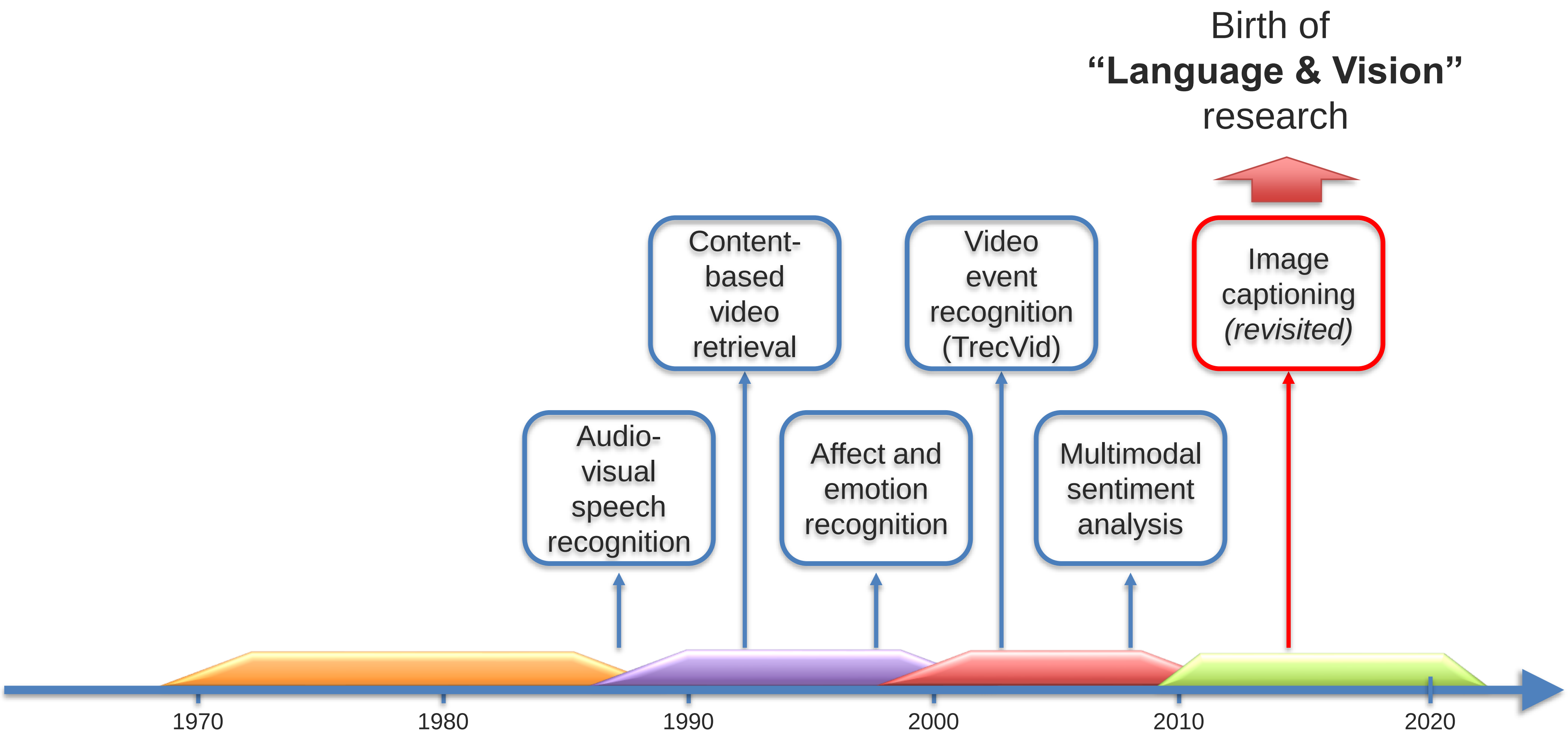
Representation learning (a.k.a. deep learning)

- 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]

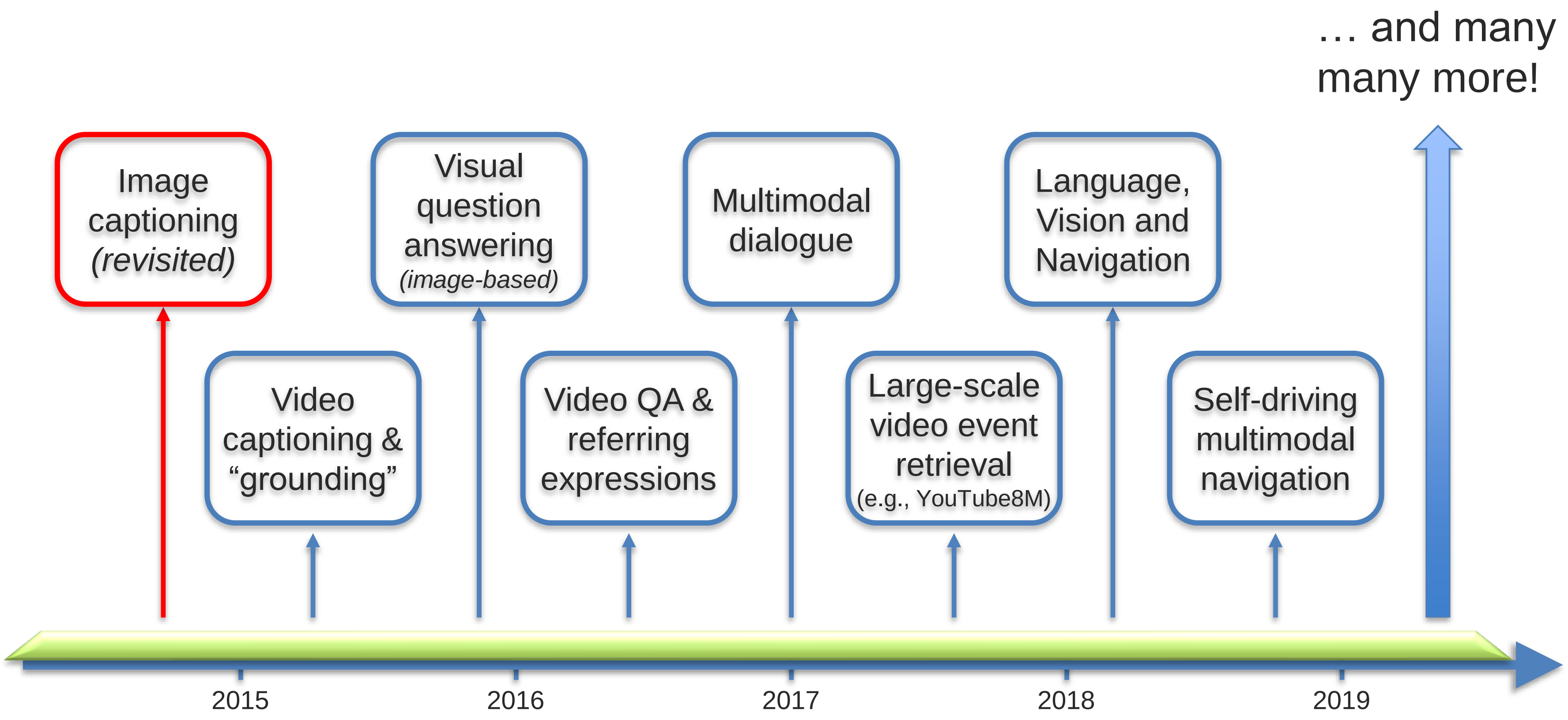
Key enablers for multimodal research:

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

Multimodal Research Tasks



Multimodal Research Tasks



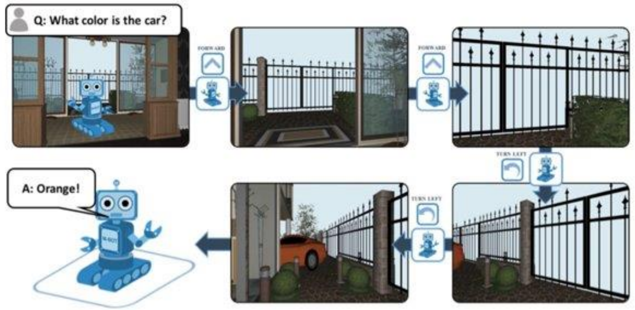
Real world tasks tackled by Multimodal ML

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

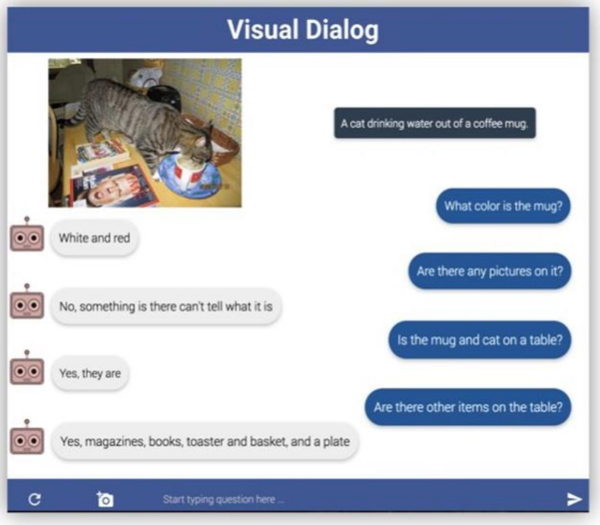


Real world tasks tackled by Multimodal ML

E. Multimodal Dialog
   - Grounded dialog

F. Event recognition
   - Action recognition
   - Segmentation

G. Multimedia information retrieval
   - Content based/Cross-media



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

Multimodal Datasets

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

Our Latest List of Multimodal Datasets

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

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

Our Latest List of Multimodal Datasets

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
Fakeddit	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!

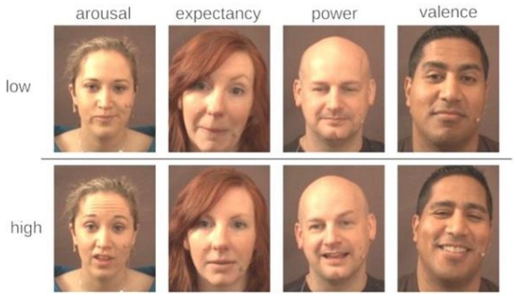
A Curated List of Multimodal Datasets

- 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 test 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



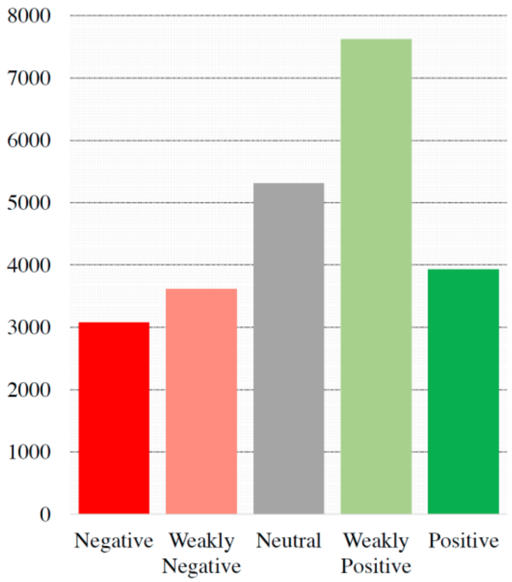
AVEC 2011/2012

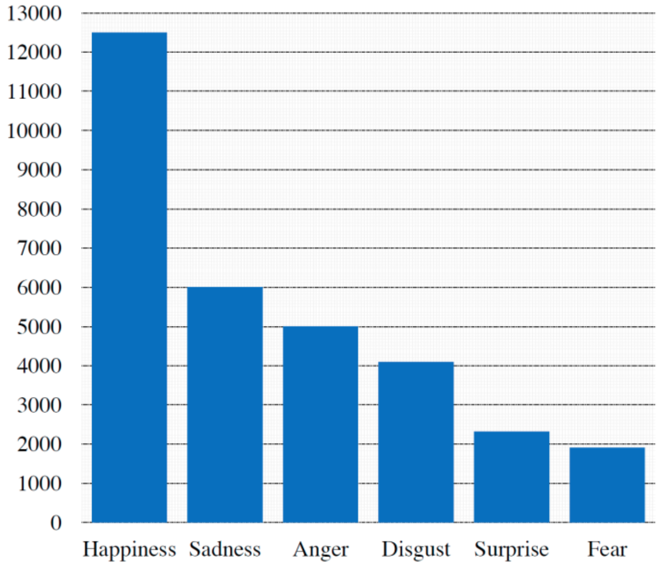
AVEC 2013/2014

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





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.

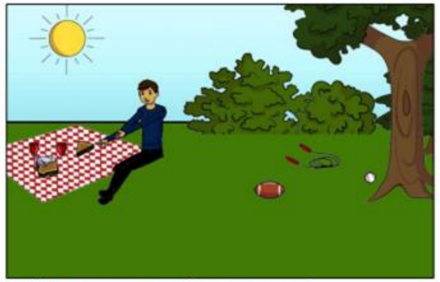
A large bus sitting next to a very tall building.

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?

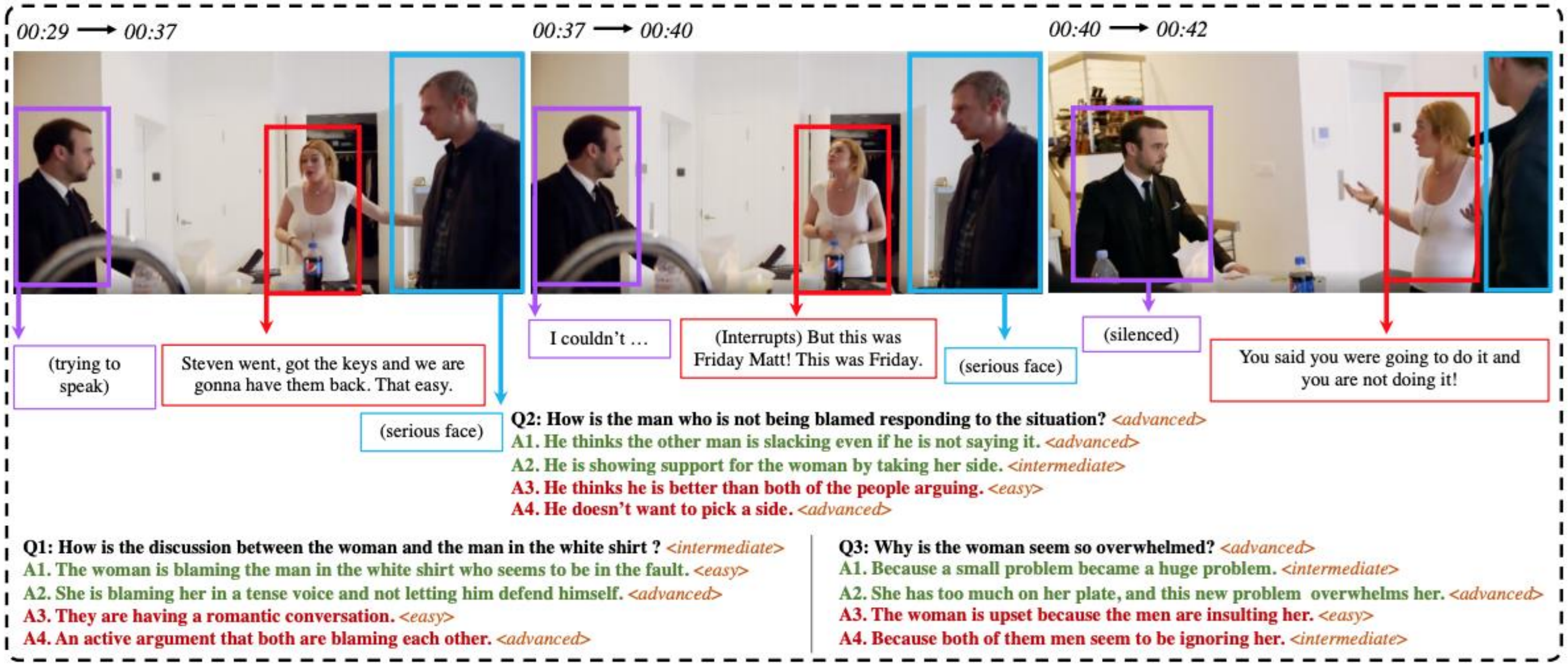


Is this person expecting company?
What is just under the tree?

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

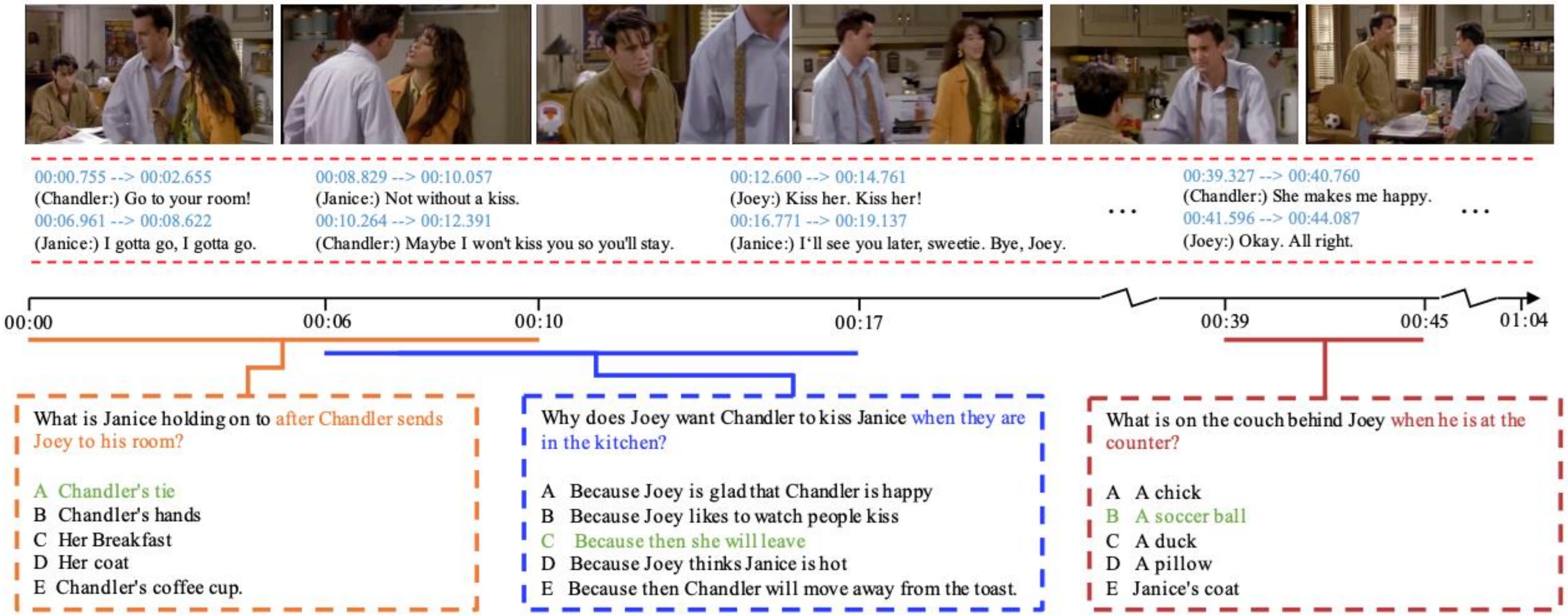
Social Interaction Q&A Dataset (A10)

- Social-IQ: 1.2k videos, 7.5k questions, 50k answers
- Questions and answers centered around social behaviors



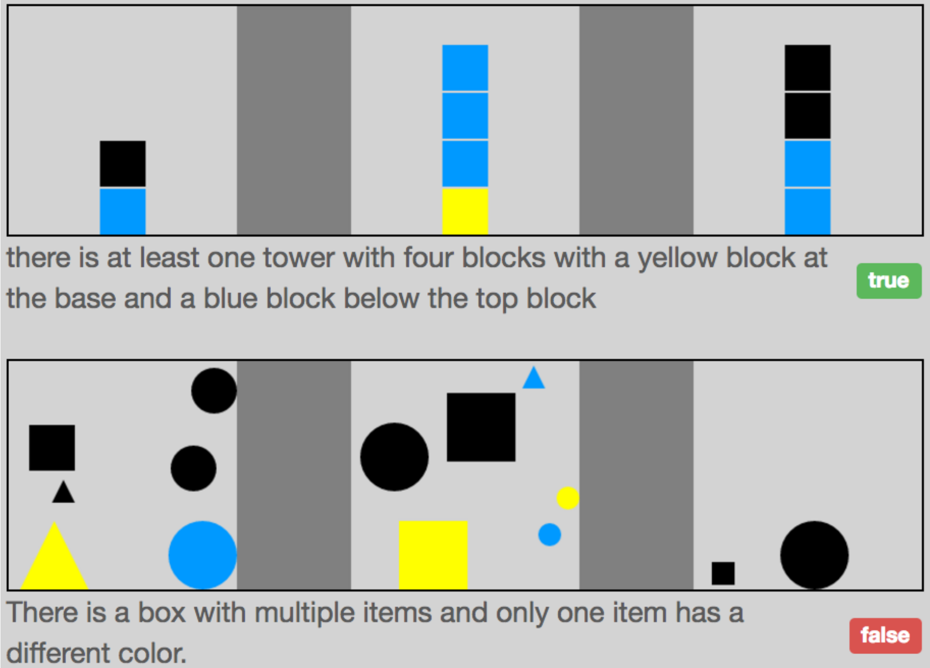
Multimodal QA (C7)

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



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

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?

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

Navigating in a Virtual House (D1)

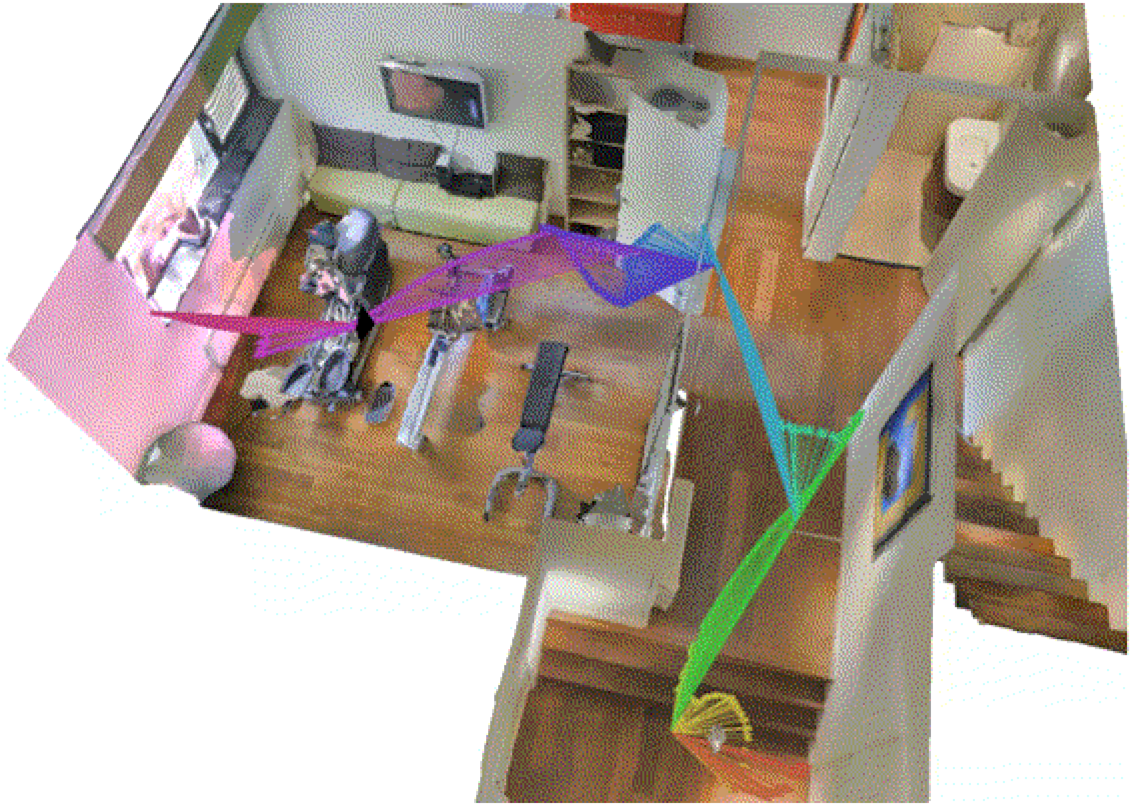
Visually-grounded natural language navigation in real buildings

- 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)

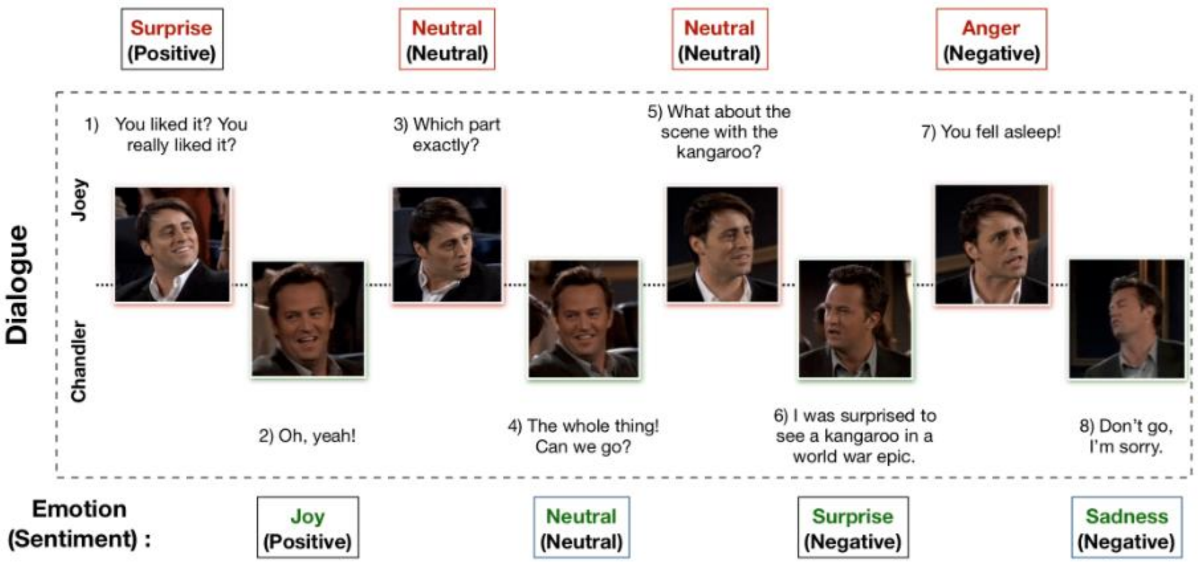
- Github
- Same words, different order, different images. Intended to test the compositionality of vision-language models

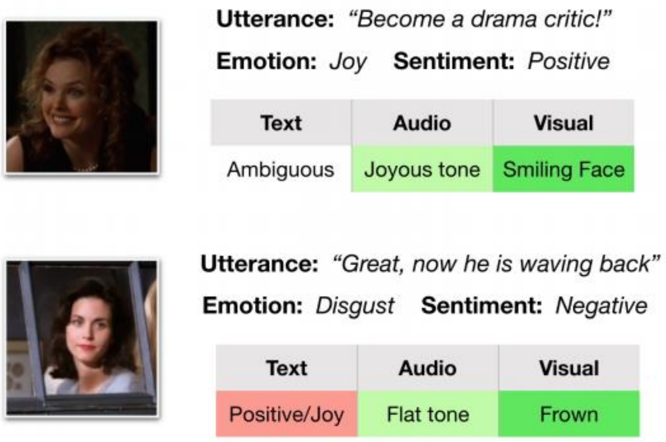
(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

Multi-Party Emotion Recognition (A11, E)

- MELD: Multi-party dataset for emotion recognition in conversations



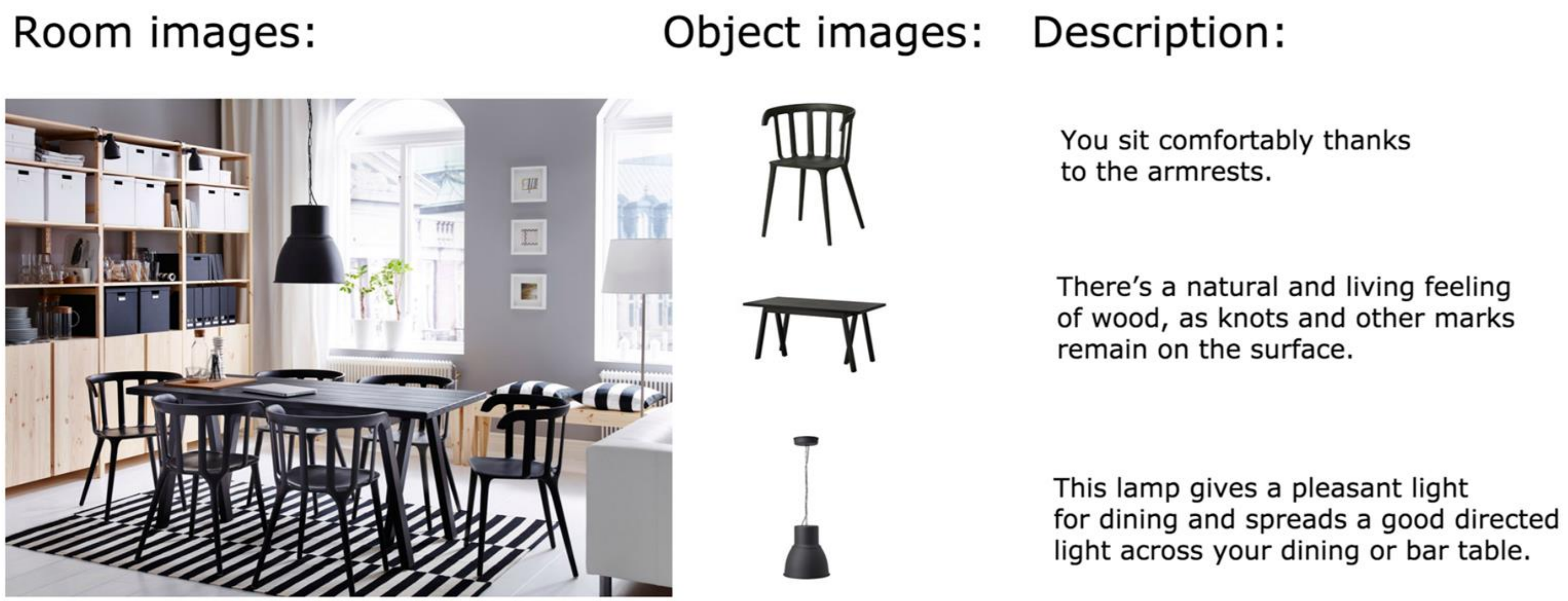


EPIC-Kitchens (F9)

- 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.



Some Advice About Multimodal Datasets

- Text, speech, audio, video
  - 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 9/5 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 9/7 to help you with finding project teammates

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

- We created the study groups in Piazza.
  - End of the discussion period: Monday 9/11 at 8pm ET

Lecture Highlights (for both lectures next week)

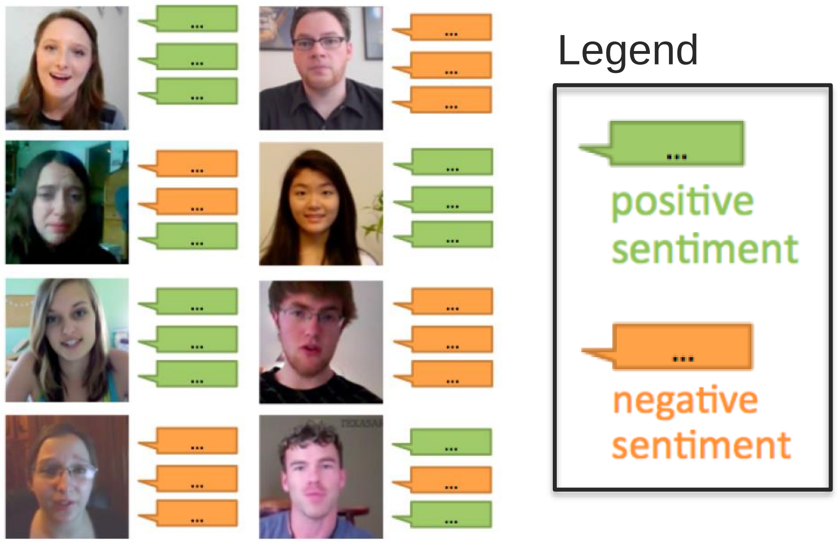
- Starting next week, you need to post your lecture highlights following each course lecture. See Piazza for detailed instructions.

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



What rules can you infer from this data?

- ✔ *Smile* -> positive sentiment
- ✔ *Frown* -> negative sentiment
- ✔ *nod*-> positive sentiment
- ✗ *Wearing glasses* -> negative sentiment

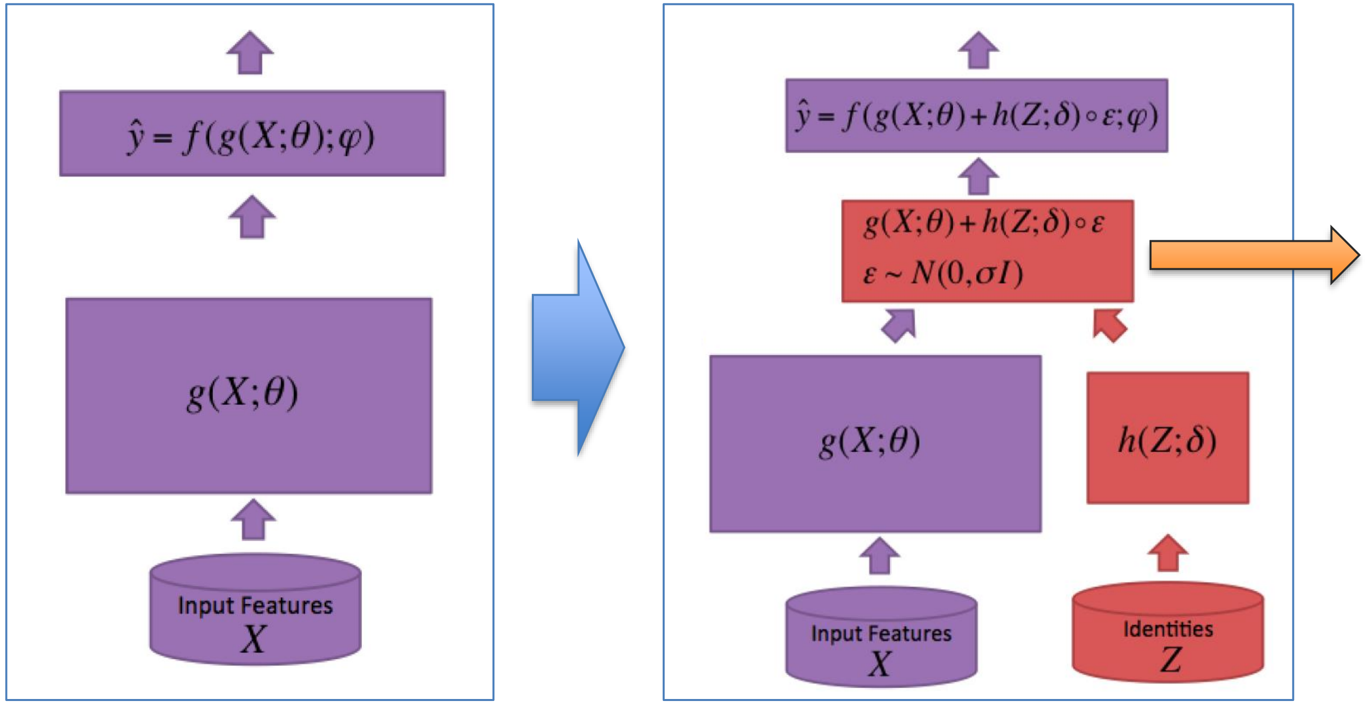
Confounding factor!

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**



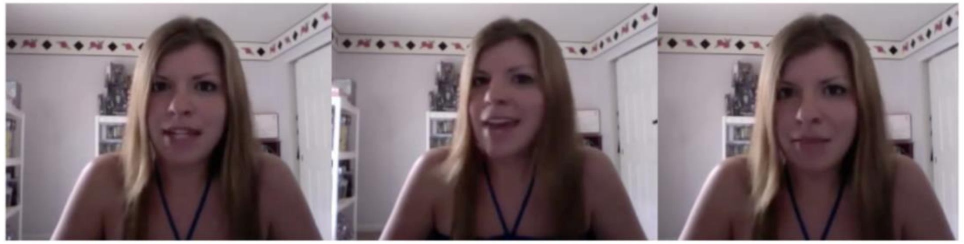
Hypothesis: the representation is a mixture from the person-independent factor g(X) and the person-dependent factor h(Z).

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

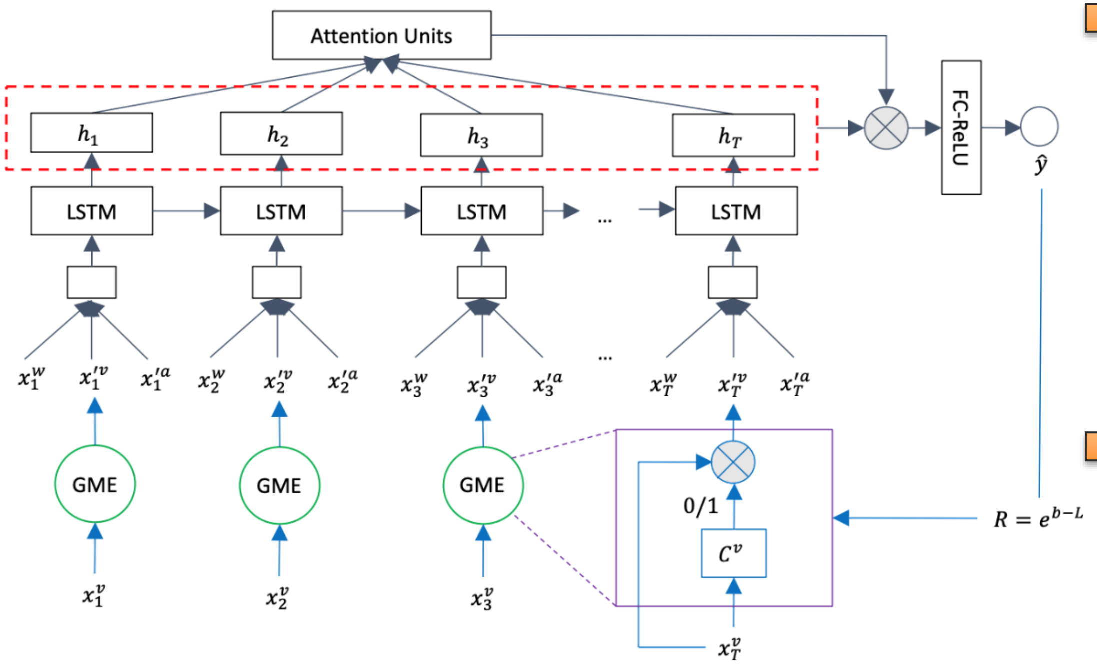
Visual modality: Hands cover mouth

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

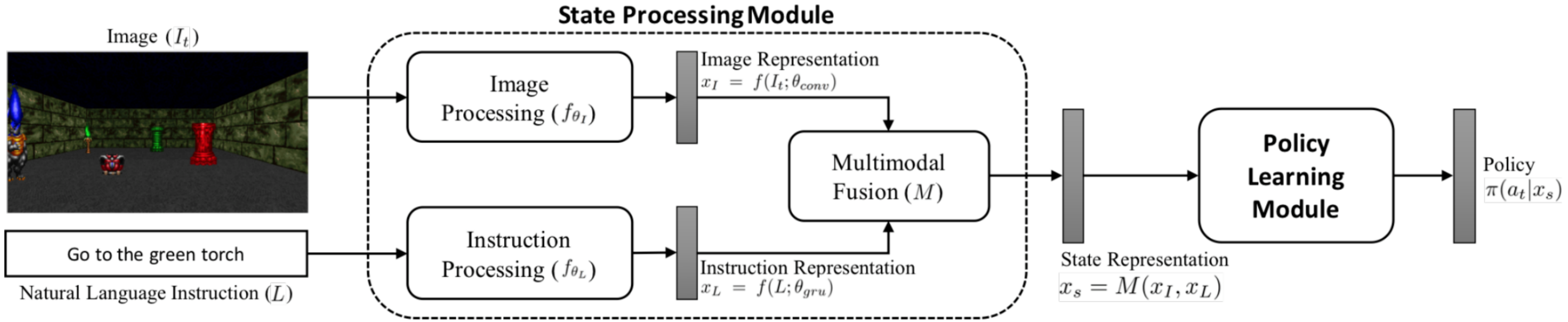
Project Example: Instruction Following

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

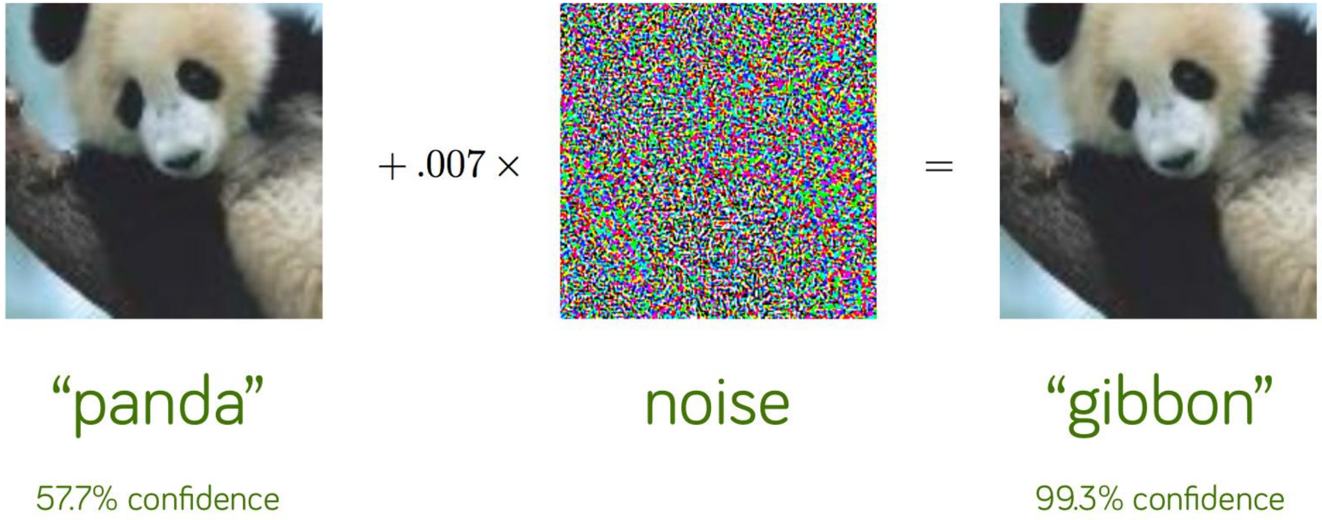


Hypothesis: Gated attention learns to ground and compose attributes in natural language with the image features. e.g. learning grounded representations for 'green' and 'torch'.

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.



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

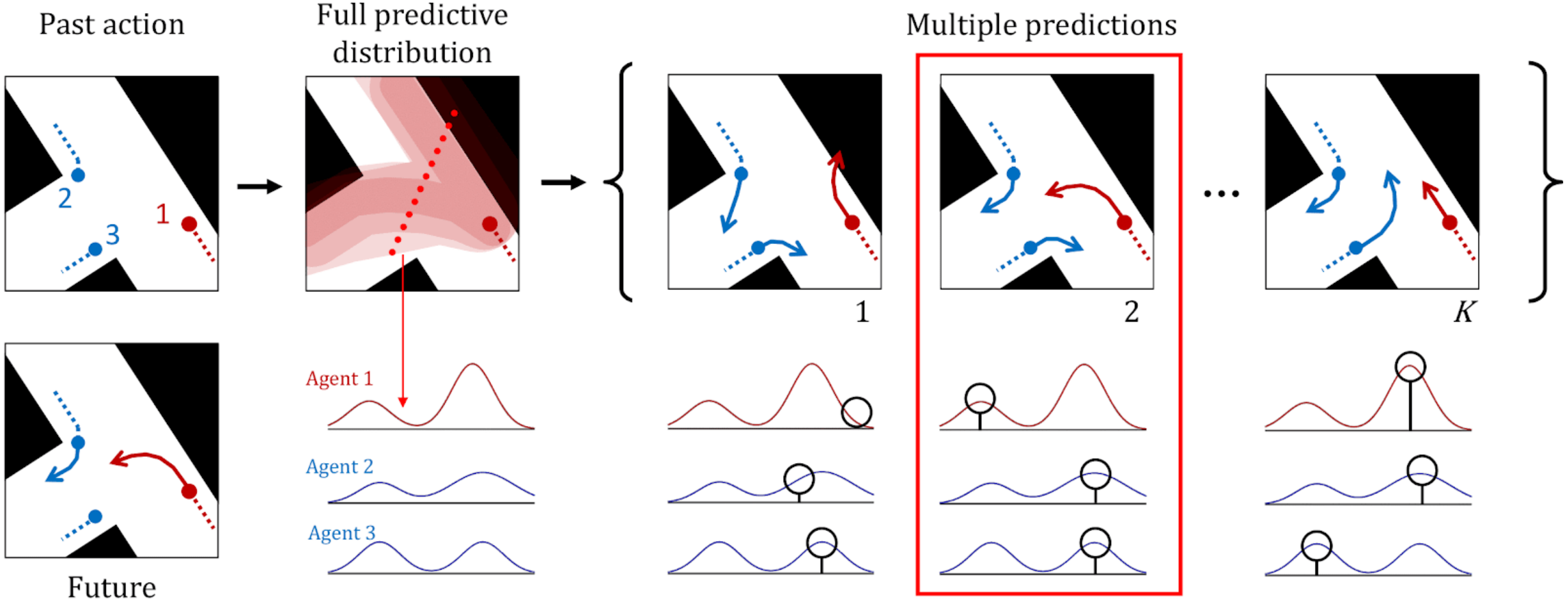
A: Roses to Sunflower

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

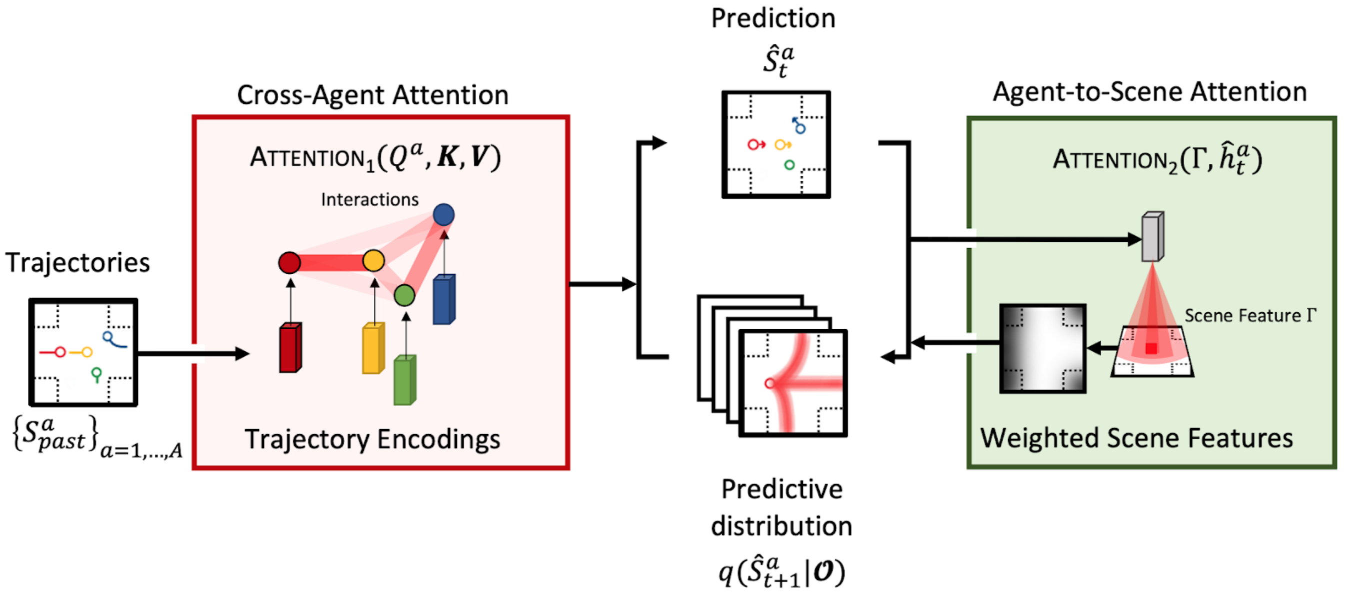
Research task: Multiagent trajectory forecasting for autonomous driving
Datasets: Argoverse and Nuscenes autonomous driving datasets
Main idea: Build a model that understands the environment and multiagent trajectories and predicts a set of multimodal future trajectories for each agent.



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.



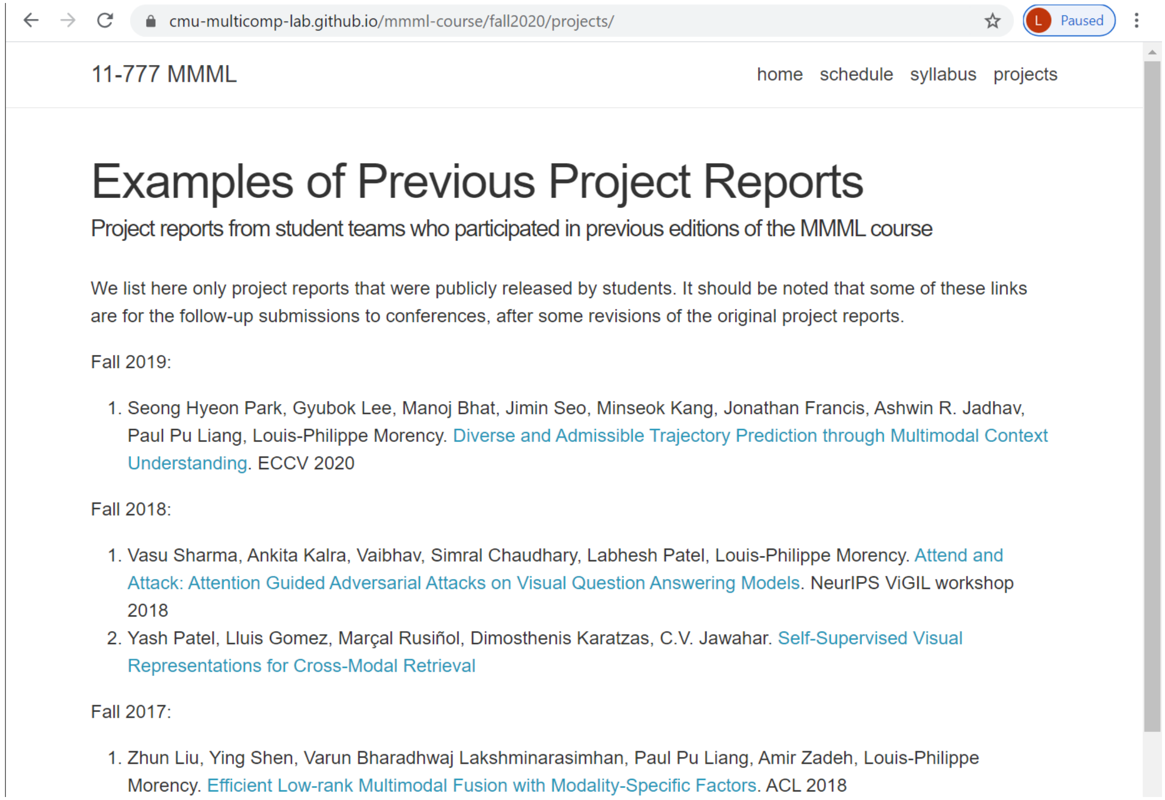
Hypothesis: both agent-agent interactions and agent-scene interactions are important!

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/



cmu-multicomp-lab.github.io/mmml-course/fall2020/projects/

Paused

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, Lluis 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