Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 1.2: Multimodal Research Tasks

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Upcoming Course Assignments

Project preferences (deadline Tuesday 9/1 at 8pm ET)

- Let us know about your project preferences, including datasets, research topics and potential teammates
  - Preferences will be shared with other students
- We will reserve a moment for discussions on Thursday 9/3 to help you with finding project teammates

Reading Assignment (due Friday 9/4 at 8pm ET)

- Instructions shared via Canvas

Lecture Participation (starting next week)

- Starting next week, please bring your cellphone to class.
  - We plan to use Poll Everywhere for class participation.

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 1.2: Multimodal Research Tasks

Louis-Philippe Morency

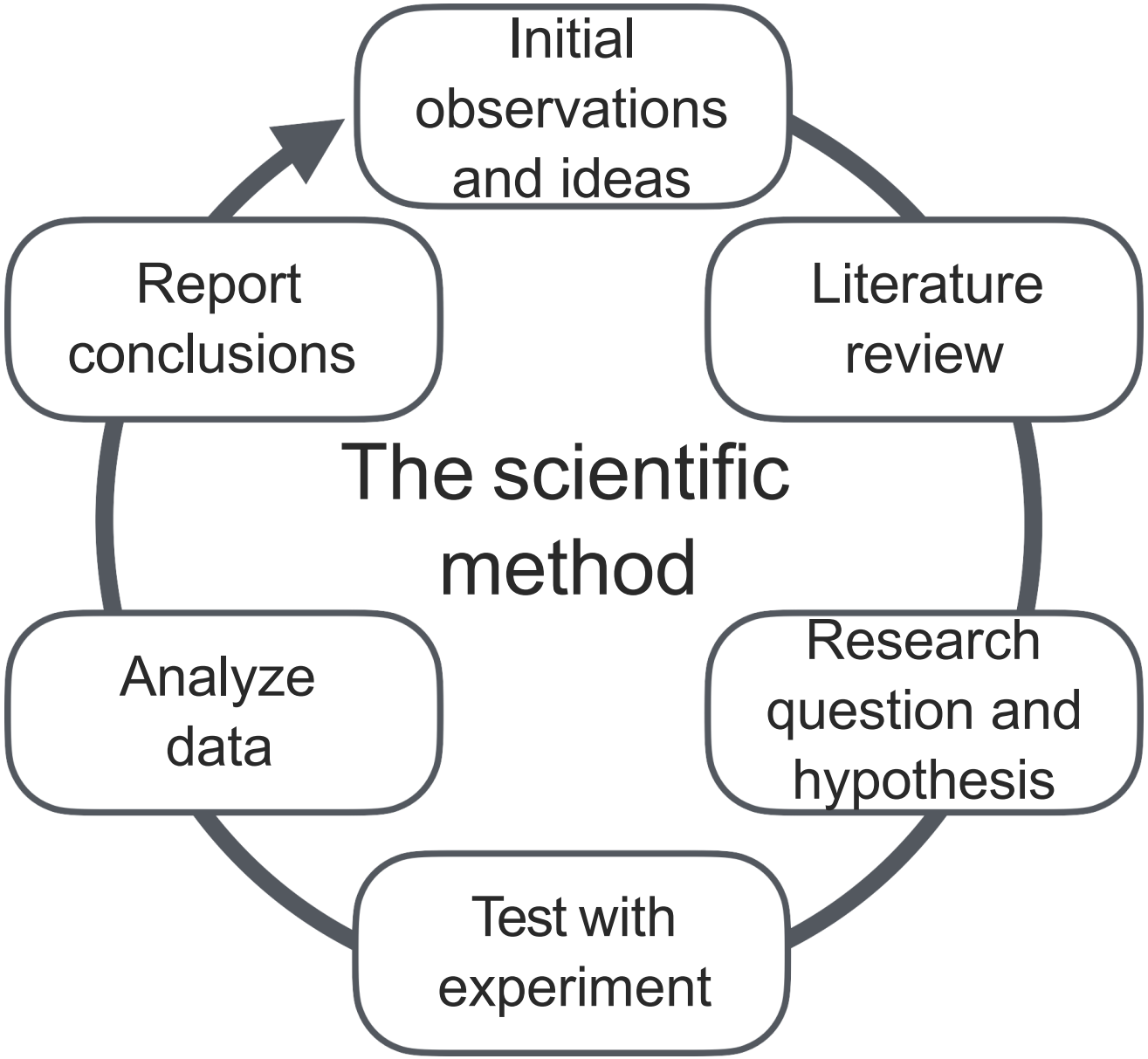
** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Lecture Objectives

- Experimental design
  - Research questions and hypotheses
- A historical view on multimodal research
- Multimodal research tasks
  - Curated list of multimodal datasets
- Word representations
  - Distributional hypothesis
  - Learning neural representations
- Appendix: Some examples of previous course projects

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.

- Meetings with TAs these coming 2 weeks
- Project presentations with instructor 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~~

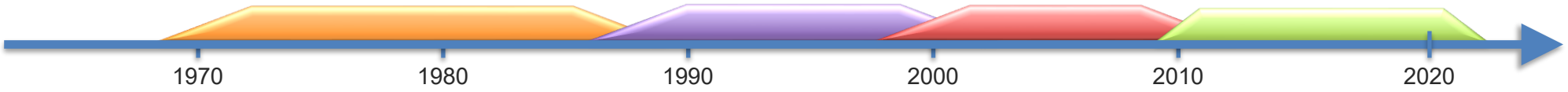
- Better understanding of the cross-modal interactions in multimodal models
- Understanding compositionality and multimodal reasoning
- Robustness to missing/noisy modalities, adversarial attacks
- Studying social biases and creating fairer models
- Interpretable and trustworthy models
- Faster and more 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

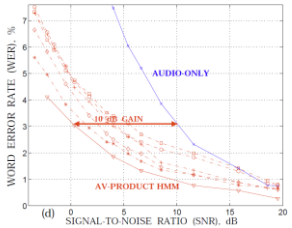
David McNeill

"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

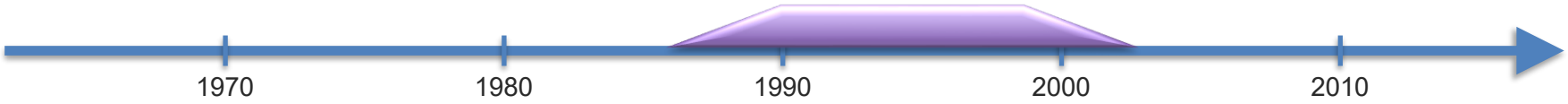
2) Multimodal interfaces

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

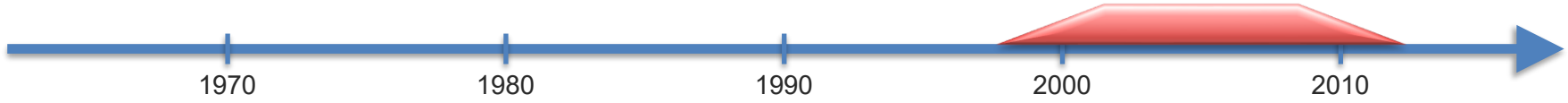
CALO Project [2003-2008, SRI]

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

Social Signal Processing Network

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)

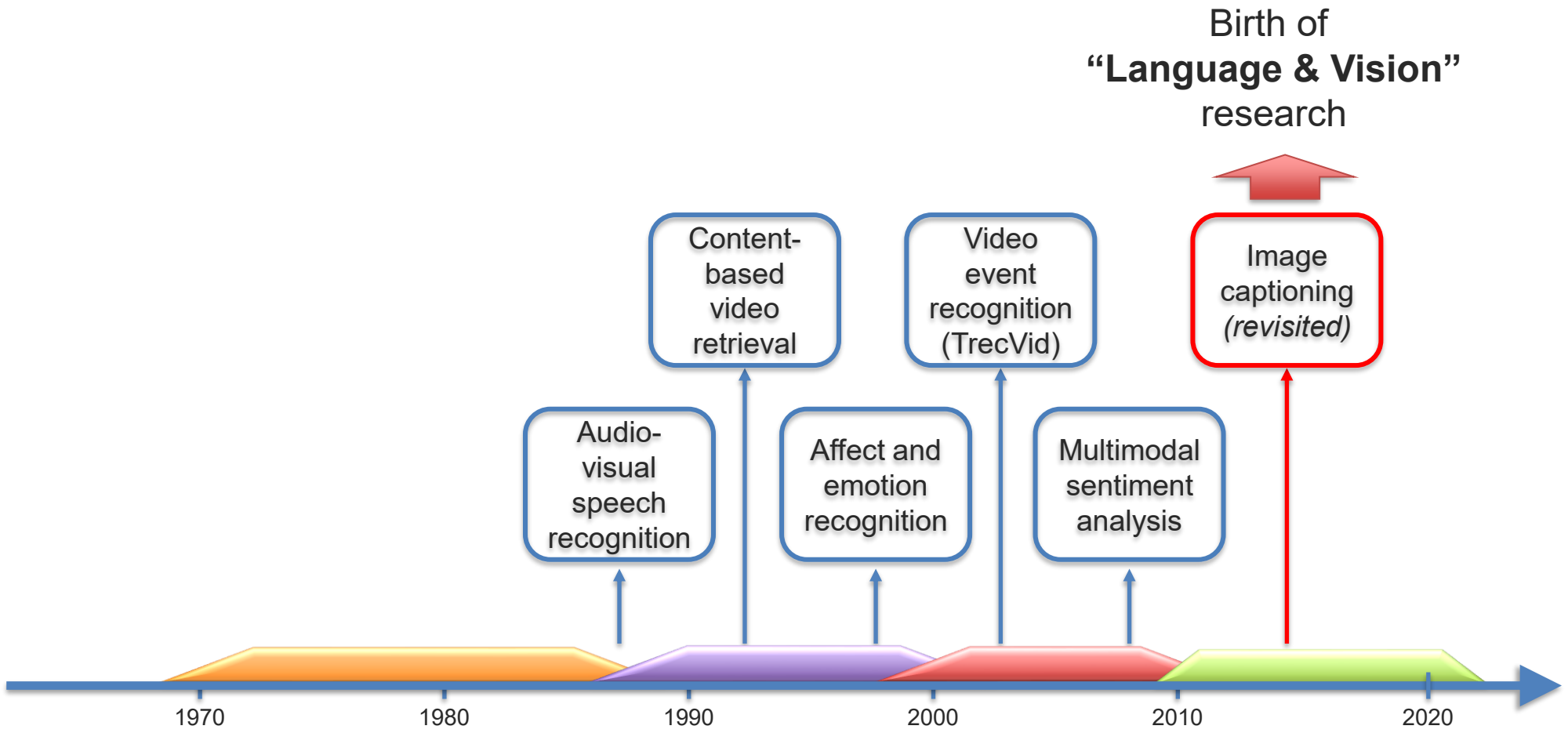
- Multimodal deep learning [ICML 2011]
- Multimodal Learning with Deep Boltzmann Machines [NIPS 2012]
- Visual attention: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [ICML 2015]

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



Multimodal Research Tasks

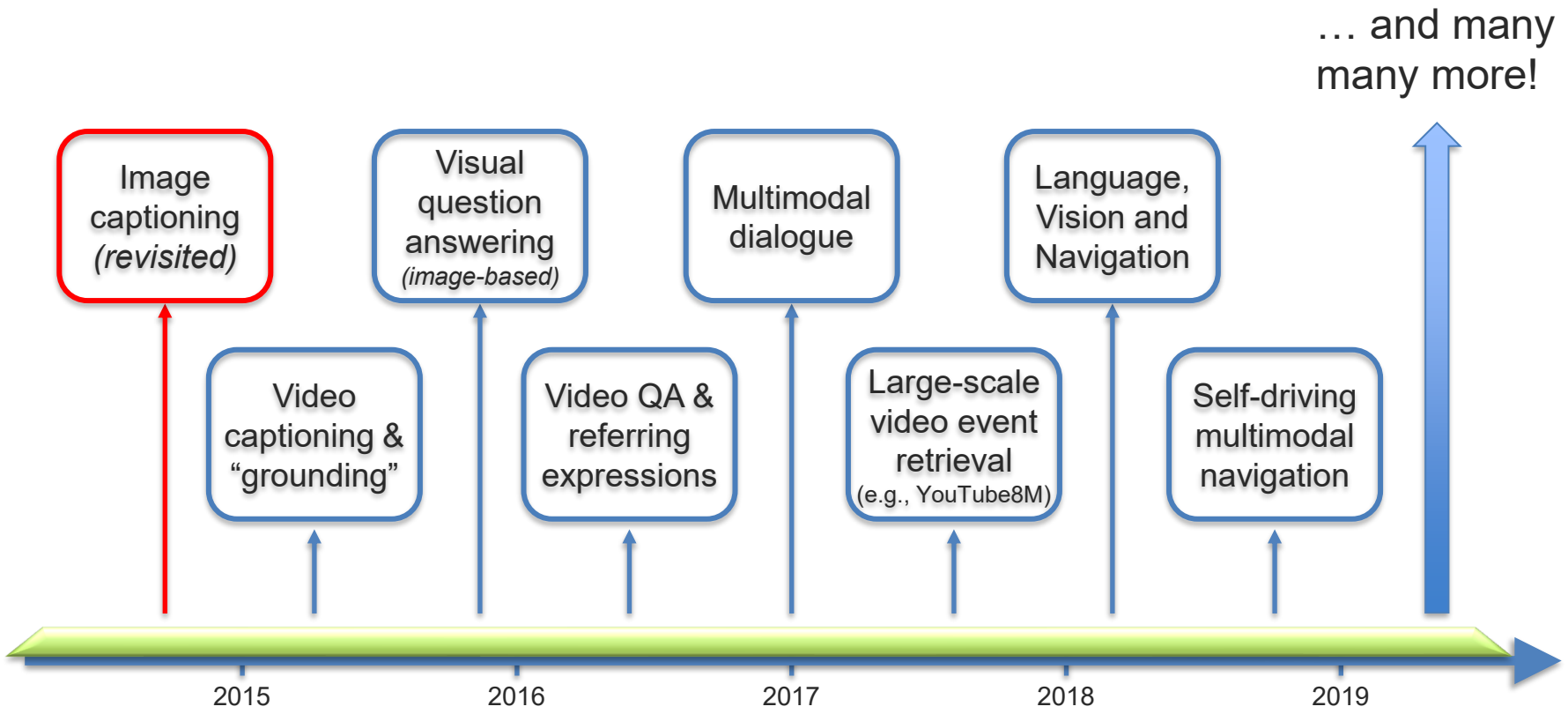


Image Captioning

- 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.

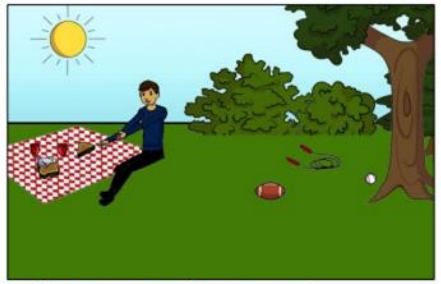
A large bus sitting next to a very tall building.

Visual Questions & Answers – VQA

- 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?

Visual Questions & Answers – MMMU

Art & Design

Question: Among the following harmonic intervals, which one is constructed incorrectly?

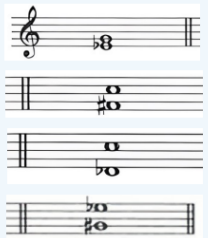
Options:

(A) Major third <*image 1*>

(B) Diminished fifth <*image 2*>

(C) Minor seventh <*image 3*>

(D) Diminished sixth <*image 4*>



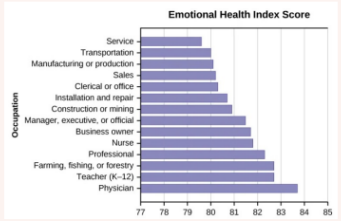
Subject: Music; **Subfield:** Music;
Image Type: Sheet Music;
Difficulty: Medium

Business

Question: ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?

Options:

(A) 0 (B) 0.2142
(C) 0.3571 (D) 0.5



Subject: Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts;
Difficulty: Medium

Science

Question: <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

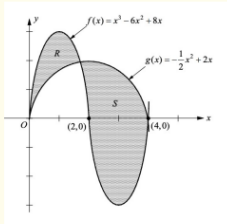
Options:

(A) $\int_0^{1.5}[f(x)-g(x)]dx$

(B) $\int_0^{1.5}[g(x)-f(x)]dx$

(C) $\int_0^{2}[f(x)-g(x)]dx$

(D) $\int_0^{2}[g(x)-x(x)]dx$



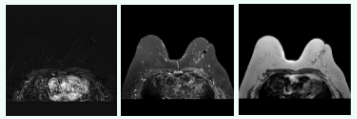
Subject: Math; **Subfield:** Calculus;
Image Type: Mathematical Notations;
Difficulty: Easy

Health & Medicine

Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast?

Options:

(A) Susceptibility artifact
(B) Hematoma
(C) Fat necrosis (D) Silicone granuloma



Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.;
Difficulty: Hard

Humanities & Social Science

Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image* 1>

Option:

(A) Oppressor
(B) Imperialist
(C) Savior (D) Isolationist

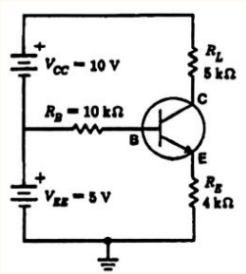
Subject: History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons;
Difficulty: Easy

Tech & Engineering

Question: Find the VCE for the circuit shown in <*image 1*>. Neglect VBE

Answer: 3.75

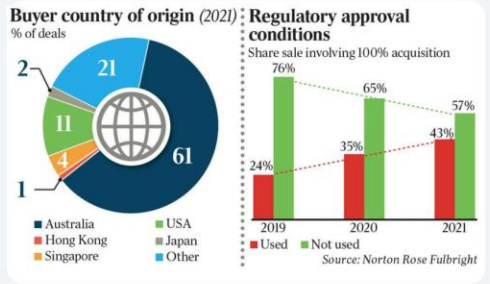
Explanation: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V



Subject: Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams;
Difficulty: Hard

Visual Questions & Answers – ChartQAPro

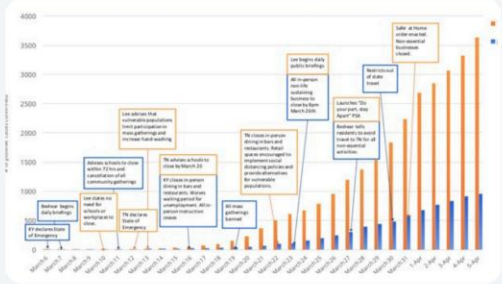
(a) Mathematical Reasoning



Question: Calculate the total percentage of deals made by buyers from the USA, Japan, and Singapore combined.

Answer: 17

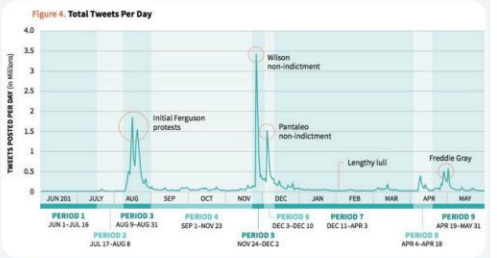
(b) Visual Reasoning



Question: At which date the blue bar had a value larger than 500 and the orange bar had a value below 2500?

Answer: March 31

(c) Conversational



Q1: How many peaks does Period 8 have?

A1: 2

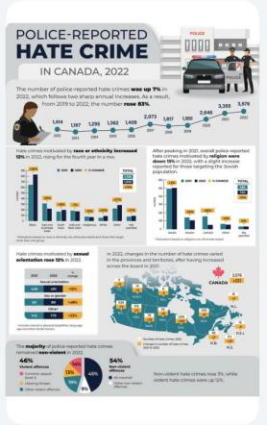
Q2: Which event caused the most significant spike in tweets per day within Period 5?

A2: Wilson non-indictment

Q3: Is this the largest peak in the graph?

A3: Yes

(d) Multiple-Choice

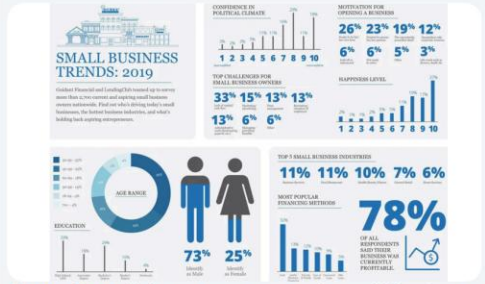


Question: What was the percentage change in hate crimes motivated by religion from 2021 to 2022?

A) 10% decrease
B) 15% decrease
C) 18% decrease
D) 20% decrease

Answer: B

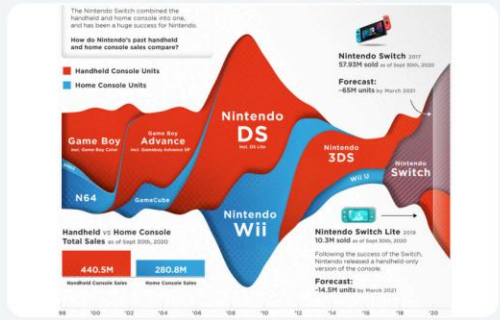
(e) Hypothetical



Question: If the percentage of small business owners identifying as male decreases by 10 percentage points, what would the new percentage be?

Answer: 63%

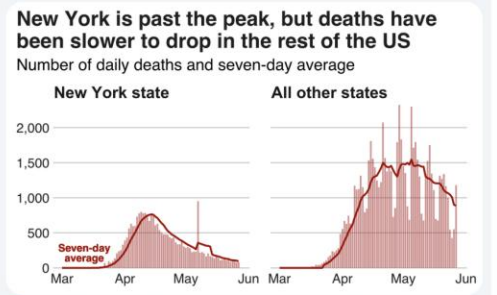
(f) Fact-Checking



Question: The Nintendo Switch Lite sold exactly one-sixth as many units as the Nintendo Switch between 2016 and 2020.

Answer: False

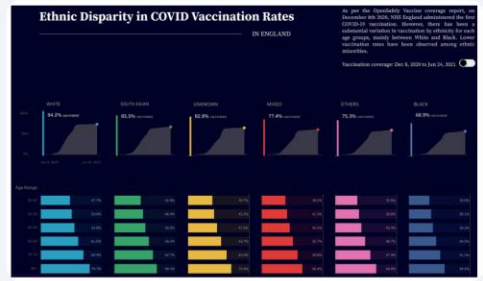
(g) Unanswerable



Question: What is the total count of hospitalizations on August 31st?

Answer: Unanswerable

(h) Multi-Chart QA

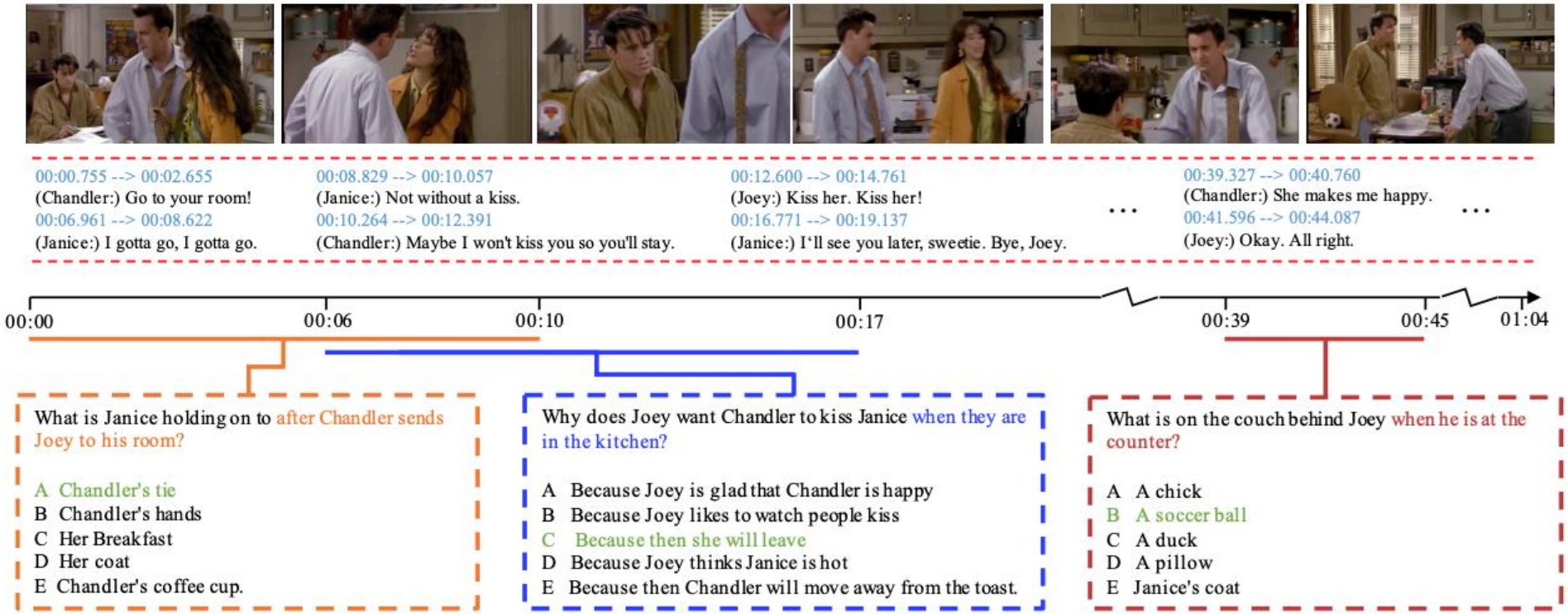


Question: What is the difference in vaccination rates between the South Asian and Mixed ethnicities in the 65-69 age group?

Answer: 5.50%

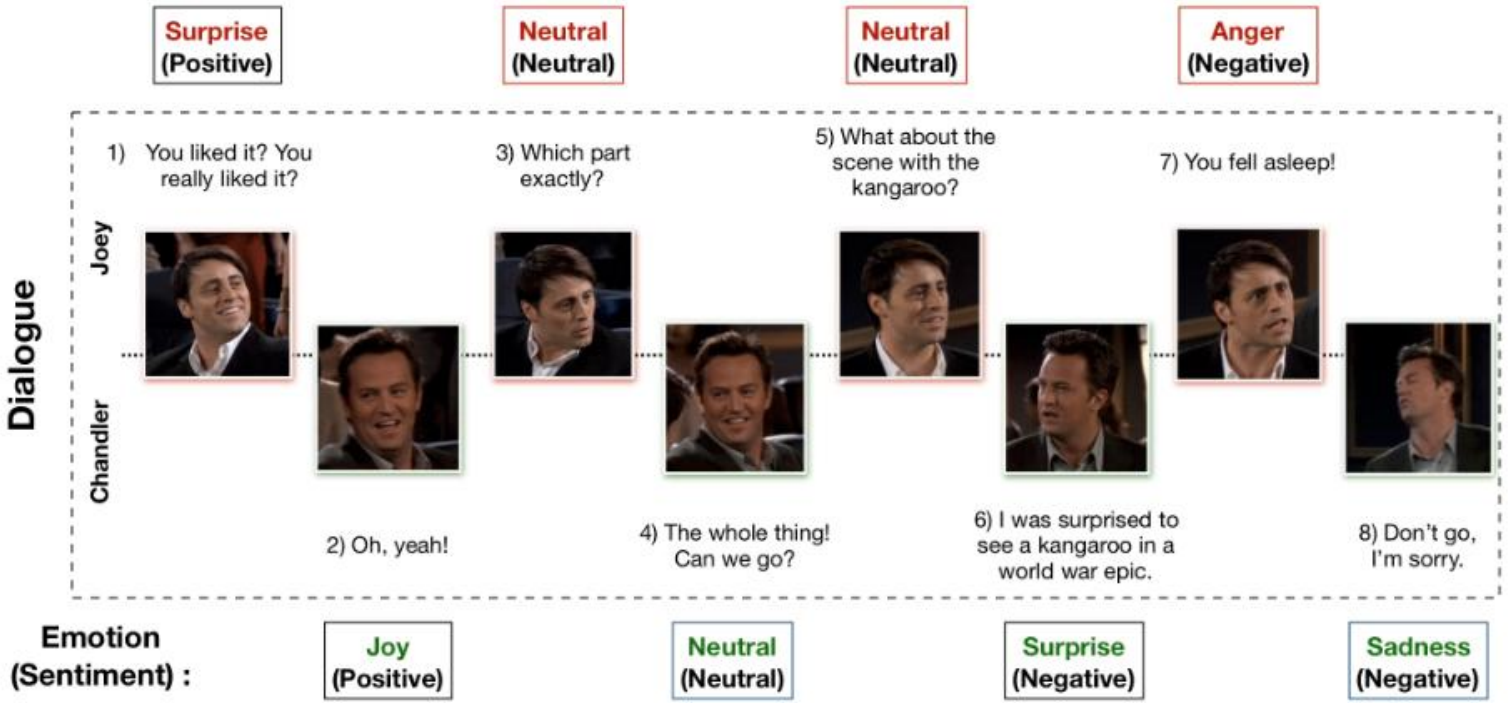
Multimodal QA

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



Multi-Party Emotion Recognition

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



Utterance: *"Become a drama critic!"*

Emotion: *Joy* **Sentiment:** *Positive*

Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face

Utterance: *"Great, now he is waving back"*

Emotion: *Disgust* **Sentiment:** *Negative*

Text	Audio	Visual
Positive/Joy	Flat tone	Frown

Navigating in a Virtual House

Visually-grounded natural language navigation in real buildings

- Room-2-Room: 21,567 open vocabulary, crowd-sourced navigation instructions

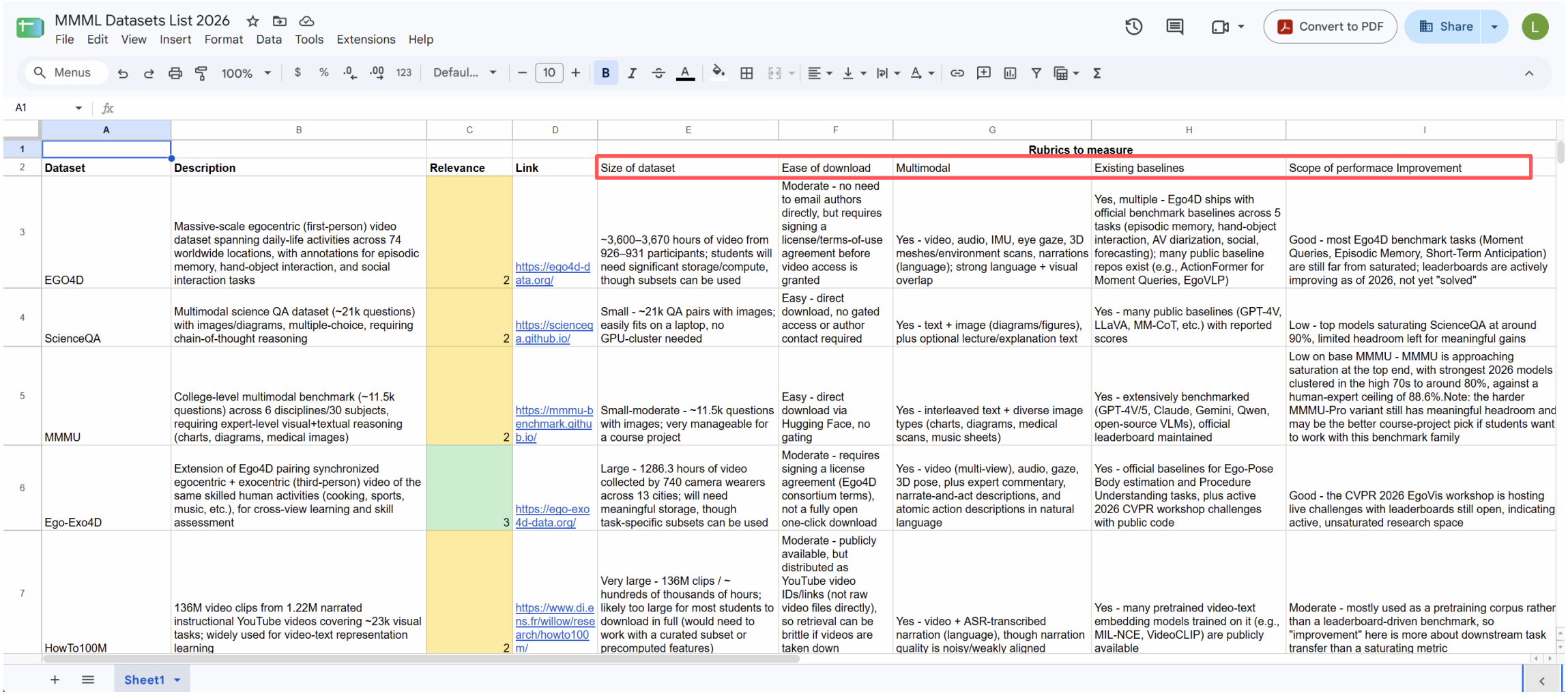
Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

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

Curated List of Multimodal Datasets

https://tinyurl.com/MMML-datasets



MMML Datasets List 2026

File Edit View Insert Format Data Tools Extensions Help

Convert to PDF | Share

	Dataset	Description	Relevance	Link	Size of dataset	Ease of download	Multimodal	Existing baselines	Scope of performace Improvement
					Rubrics to measure				
3	EGO4D	Massive-scale egocentric (first-person) video dataset spanning daily-life activities across 74 worldwide locations, with annotations for episodic memory, hand-object interaction, and social interaction tasks	2	https://ego4d-data.org/	~3,600–3,670 hours of video from 926–931 participants; students will need significant storage/compute, though subsets can be used	Moderate - no need to email authors directly, but requires signing a license/terms-of-use agreement before video access is granted	Yes - video, audio, IMU, eye gaze, 3D meshes/environment scans, narrations (language); strong language + visual overlap	Yes, multiple - Ego4D ships with official benchmark baselines across 5 tasks (episodic memory, hand-object interaction, AV diarization, social, forecasting); many public baseline repos exist (e.g., ActionFormer for Moment Queries, EgoVLP)	Good - most Ego4D benchmark tasks (Moment Queries, Episodic Memory, Short-Term Anticipation) are still far from saturated; leaderboards are actively improving as of 2026, not yet "solved"
4	ScienceQA	Multimodal science QA dataset (~21k questions) with images/diagrams, multiple-choice, requiring chain-of-thought reasoning	2	https://scienceqa.github.io/	Small - ~21k QA pairs with images; easily fits on a laptop, no GPU-cluster needed	Easy - direct download, no gated access or author contact required	Yes - text + image (diagrams/figures), plus optional lecture/explanation text	Yes - many public baselines (GPT-4V, LLaVA, MM-CoT, etc.) with reported scores	Low - top models saturating ScienceQA at around 90%, limited headroom left for meaningful gains
5	MMMU	College-level multimodal benchmark (~11.5k questions) across 6 disciplines/30 subjects, requiring expert-level visual+textual reasoning (charts, diagrams, medical images)	2	https://mmmu-benchmark.github.io/	Small-moderate - ~11.5k questions with images; very manageable for a course project	Easy - direct download via Hugging Face, no gating	Yes - interleaved text + diverse image types (charts, diagrams, medical scans, music sheets)	Yes - extensively benchmarked (GPT-4V/5, Claude, Gemini, Qwen, open-source VLMs), official leaderboard maintained	Low on base MMMU - MMMU is approaching saturation at the top end, with strongest 2026 models clustered in the high 70s to around 80%, against a human-expert ceiling of 88.6%.Note: the harder MMMU-Pro variant still has meaningful headroom and may be the better course-project pick if students want to work with this benchmark family
6	Ego-Exo4D	Extension of Ego4D pairing synchronized egocentric + exocentric (third-person) video of the same skilled human activities (cooking, sports, music, etc.), for cross-view learning and skill assessment	3	https://ego-exo4d-data.org/	Large - 1286.3 hours of video collected by 740 camera wearers across 13 cities; will need meaningful storage, though task-specific subsets can be used	Moderate - requires signing a license agreement (Ego4D consortium terms), not a fully open one-click download	Yes - video (multi-view), audio, gaze, 3D pose, plus expert commentary, narrate-and-act descriptions, and atomic action descriptions in natural language	Yes - official baselines for Ego-Pose Body estimation and Procedure Understanding tasks, plus active 2026 CVPR workshop challenges with public code	Good - the CVPR 2026 EgoVis workshop is hosting live challenges with leaderboards still open, indicating active, unsaturated research space
7	HowTo100M	136M video clips from 1.22M narrated instructional YouTube videos covering ~23k visual tasks; widely used for video-text representation learning	2	https://www.di.ens.fr/willow/research/howto100m/	Very large - 136M clips / ~ hundreds of thousands of hours; likely too large for most students to download in full (would need to work with a curated subset or precomputed features)	Moderate - publicly available, but distributed as YouTube video IDs/links (not raw video files directly), so retrieval can be brittle if videos are taken down	Yes - video + ASR-transcribed narration (language), though narration quality is noisy/weakly aligned	Yes - many pretrained video-text embedding models trained on it (e.g., MIL-NCE, VideoCLIP) are publicly available	Moderate - mostly used as a pretraining corpus rather than a leaderboard-driven benchmark, so "improvement" here is more about downstream task transfer than a saturating metric

Sheet1

NOTE: We also included an old list of multimodal datasets, from Fall 2023 (in Canvas)

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
- AI agents can require a lot of computation – be mindful!
- Plan accordingly!
  - Sometimes tricky to experiment on a laptop (might need to do it on a subset of data)

Available Tools

- We will be getting AWS credit: probably $150 / student
  - We will send out a request for your AWS IDs next week.
- Google Colab for small-scale experimentation
- Request will be made for LiteLLM credits
- Use available tools in your research groups
  - Or pair up with someone that has access to them

Word Representations

Unimodal Representation – Language Modality

Written language

★★★★★ **Masterful!**

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

Spoken language

MARTHA(CON'T)
Look around you. Look at all the great things you've done and the people you've helped.

CLARK
But you've only put up the good things they say about me.

MARTHA
Clark, honey. If I were to use the bad things they say I could cover the barn, the house and the outhouse.

Input observation x_i

"one-hot" vector

$|x_i|$ = number of words in dictionary

Word-level classification

Part-of-speech ?
(noun, verb,…)

Sentiment ?
(positive or negative)

Named entity ?
(names of person,…)

What is the meaning of “bardiwac”?

- He handed her her glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feet, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia’s sunshine.
- I dined off bread and cheese and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

⇒ bardiwac is a heavy red alcoholic beverage made from grapes

How to learn (word) features/representations?

Distribution hypothesis: Approximate the word meaning by its surrounding words

Words used in a similar context will lie close together

Geometric interpretation

- row vector $\mathbf{x}_{dog}$ describes usage of word *dog* in the corpus

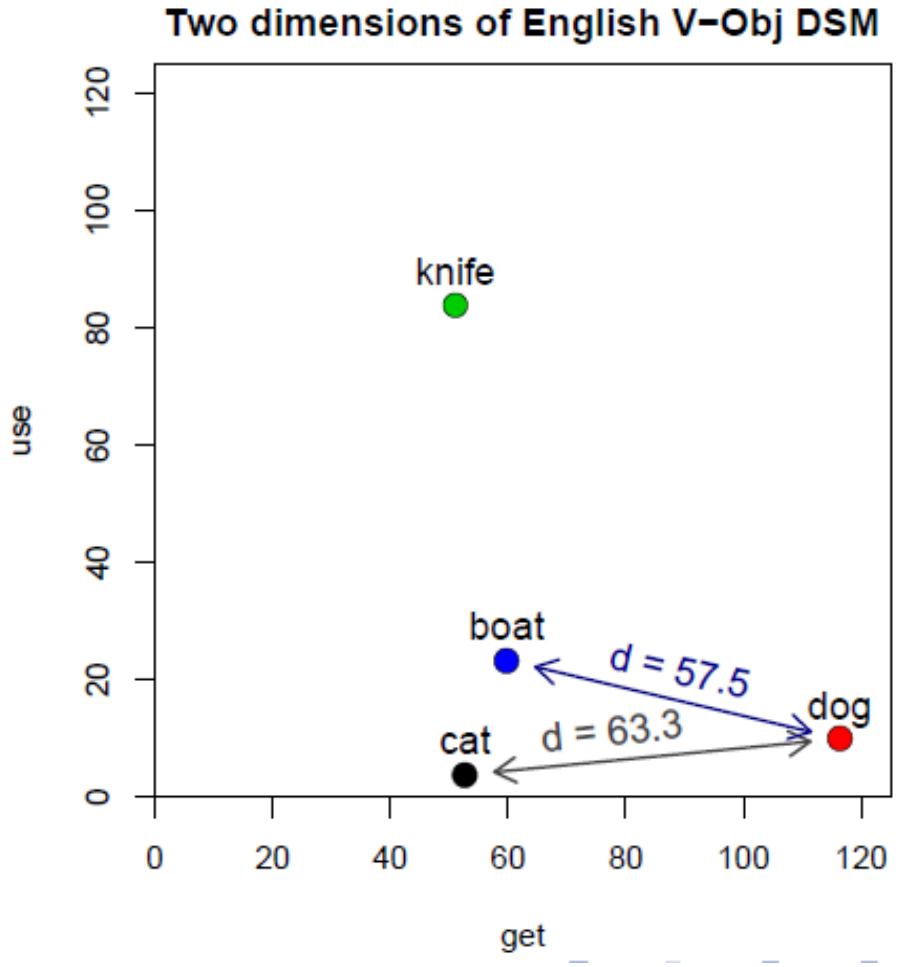
- can be seen as coordinates of point in *n*-dimensional Euclidean space R^n

	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

co-occurrence matrix M

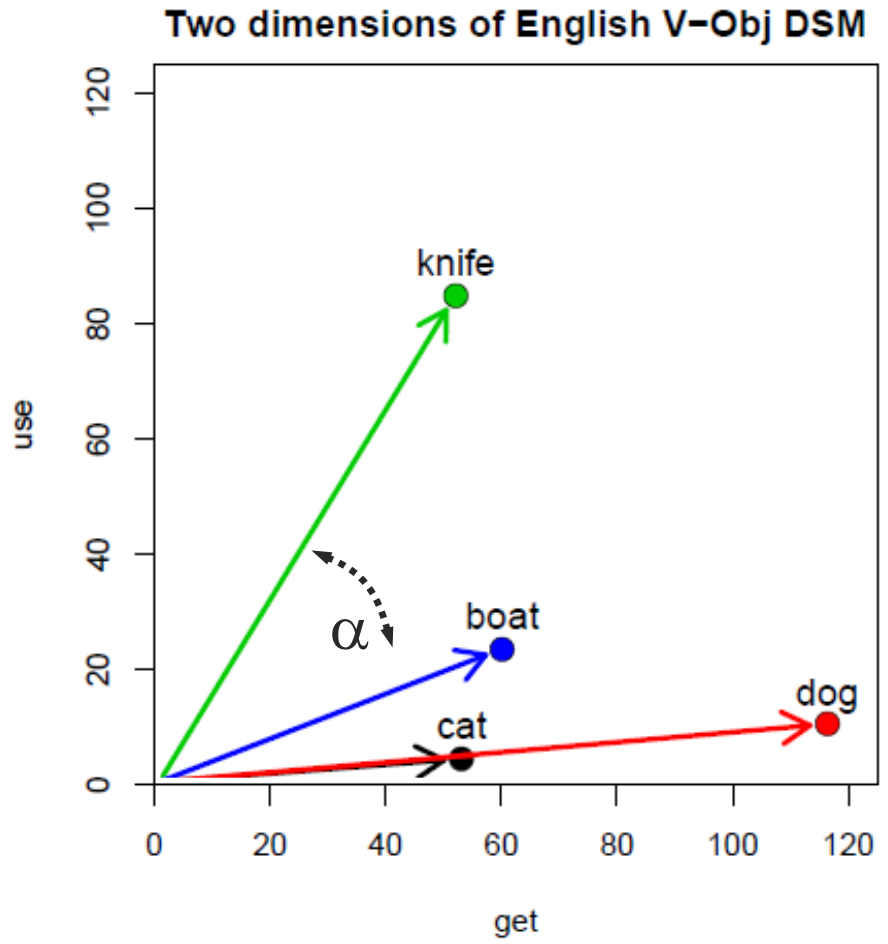
Distance and similarity

- illustrated for two dimensions: *get* and *use*: $\mathbf{x}_{dog}$ = (115, 10)
- similarity = spatial proximity (Euclidean distance)
- location depends on frequency of noun ($f_{dog} \approx 2.7 \cdot f_{cat}$)



Angle and similarity

- direction more important than location
- normalise "length" $||\mathbf{x}_{dog}||$ of vector
- or use angle α as distance measure



How to learn (word) features/representations?

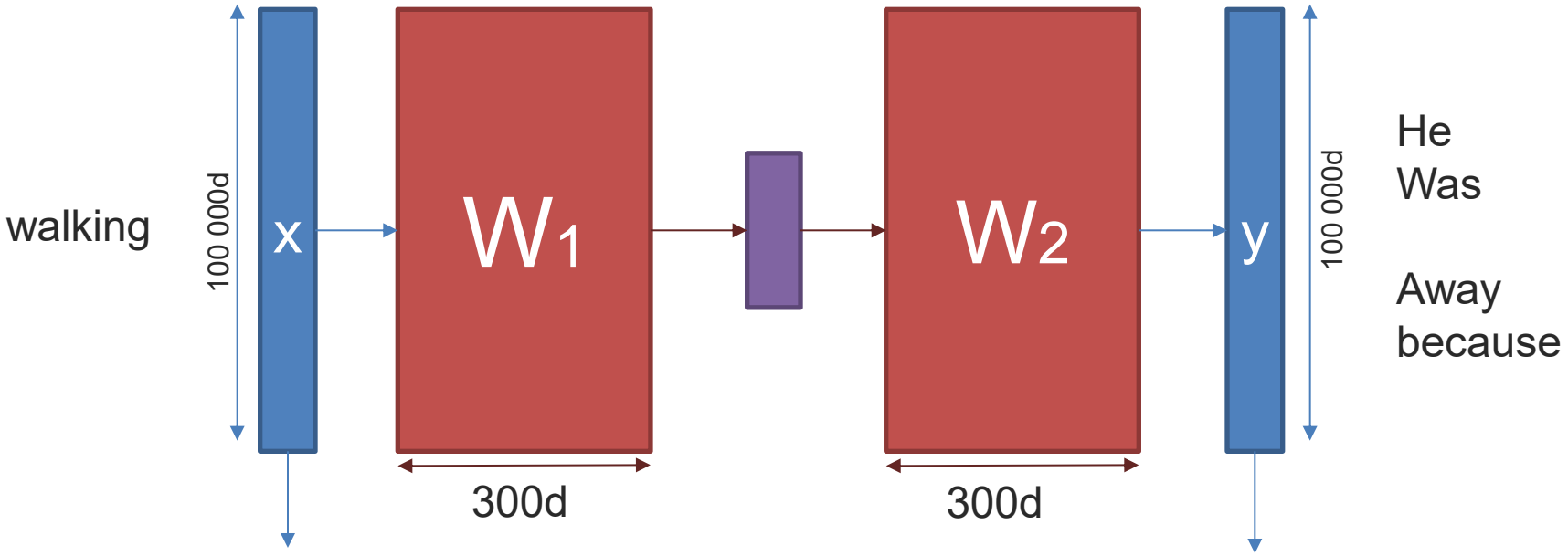
Distribution hypothesis: Approximate the word meaning by its surrounding words

Words used in a similar context will lie close together

Instead of capturing co-occurrence counts directly, predict surrounding words of every word

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

How to learn (word) features/representations?



Word2vec algorithm: https://code.google.com/p/word2vec/

How to use these word representations

If we would have a vocabulary of 100 000 words:

Classic NLP: 100 000 dimensional vector

Walking: [0; 0; 0; 0;….; 0; 0; 1; 0;…; 0; 0]

Running: [0; 0; 0; 0;….; 0; 0; 0; 0;…; 1; 0]

Similarity = 0.0

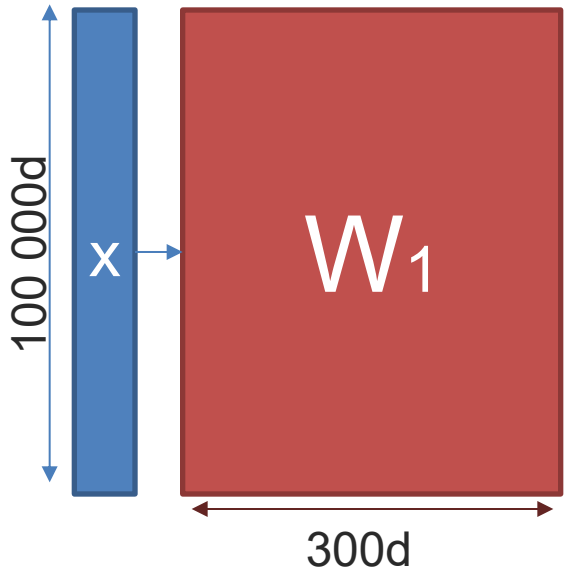
Transform: x'=x*W

Goal: 300 dimensional vector

Walking: [0,1; 0,0003; 0;….; 0,02; 0.08; 0,05]

Running: [0,1; 0,0004; 0;….; 0,01; 0.09; 0,05]

Similarity = 0.9

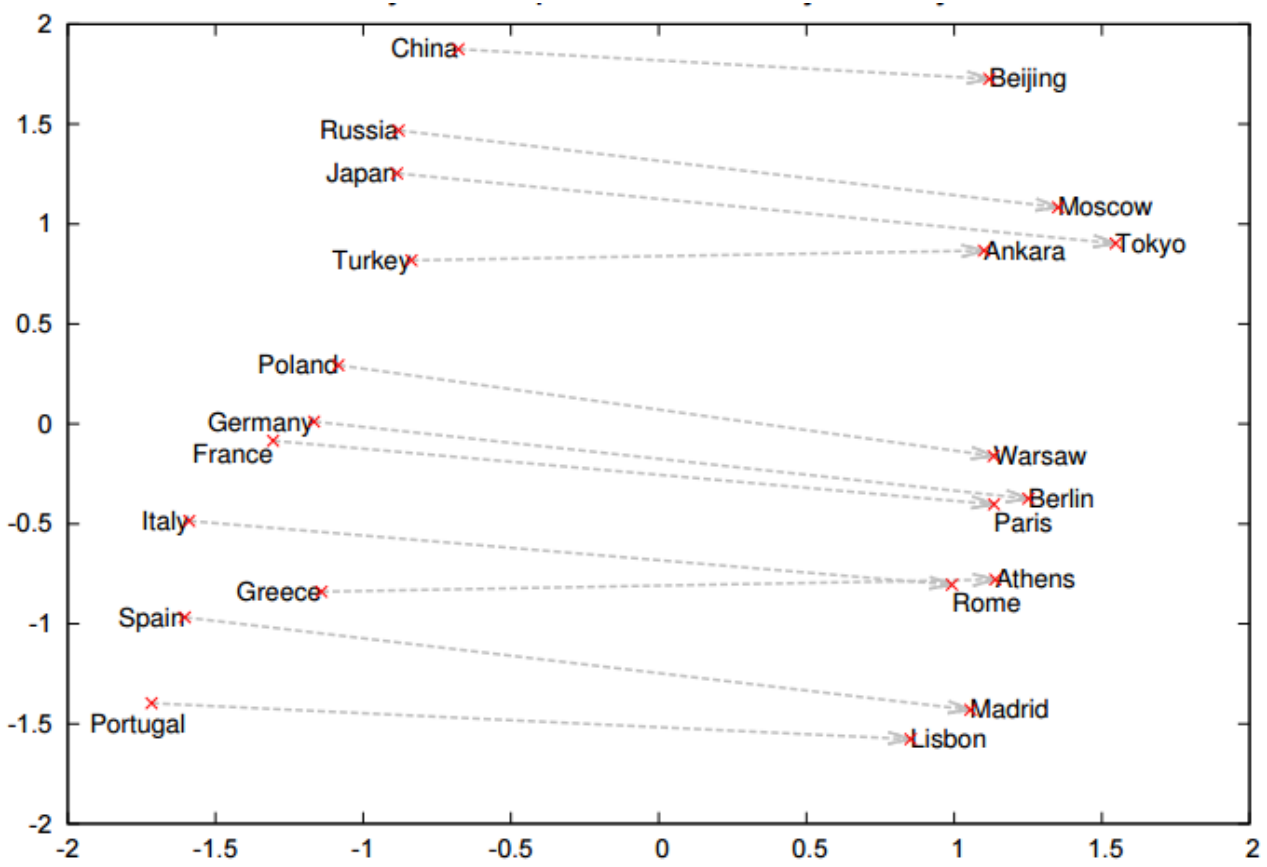


Vector space models of words

- While learning these word representations, we are actually building a vector space in which all words reside with certain relationships between them
- Encodes both syntactic and semantic relationships
- This vector space allows for algebraic operations:

Vec(king) – vec(man) + vec(woman) ≈ vec(queen)

Vector space models of words: semantic relationships



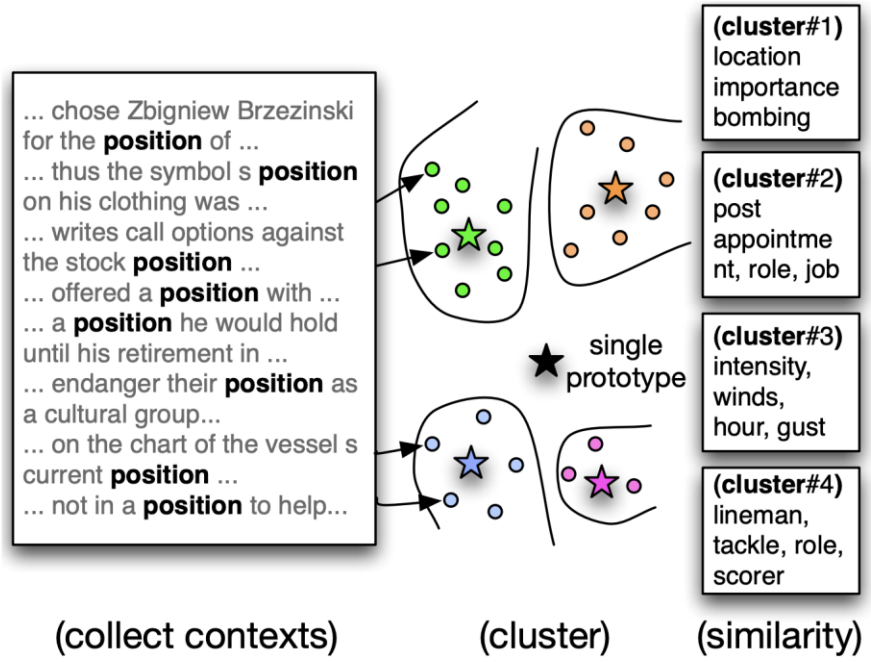
Trained on the Google news corpus with over 300 billion words

Do these work?
Issues of bias here

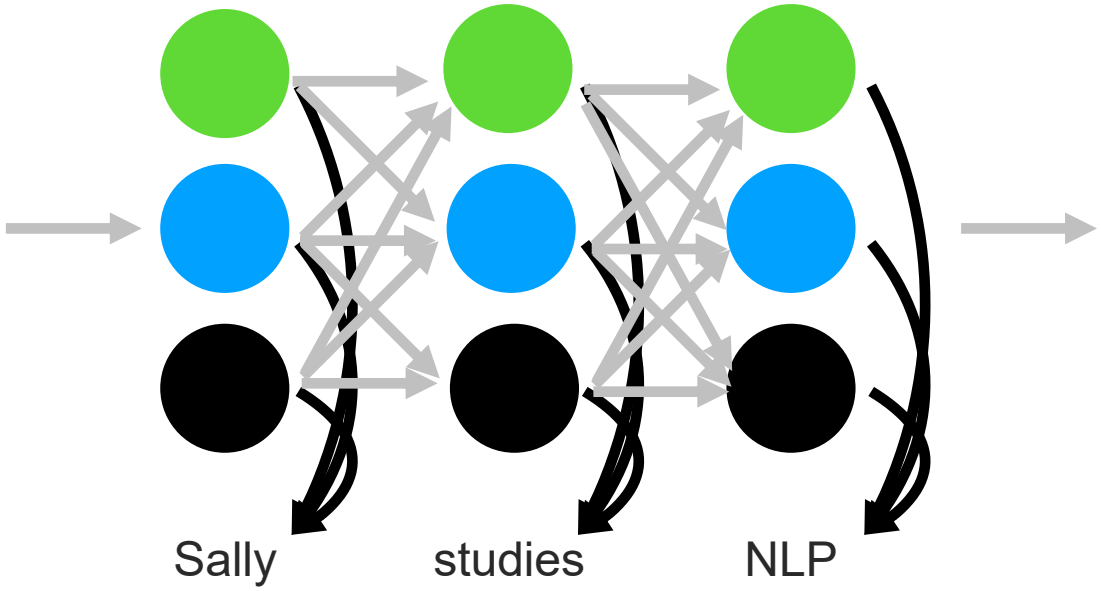
e.g. https://arxiv.org/abs/1607.06520
https://aclanthology.org/W14-1618.pdf

vec(programmer) – vec(man) +
vec(woman) ≈ vec(homemaker)

Context & Representational Power



[Reisinger and Mooney, 2010]



Senses: Beyond one-vector-per-word

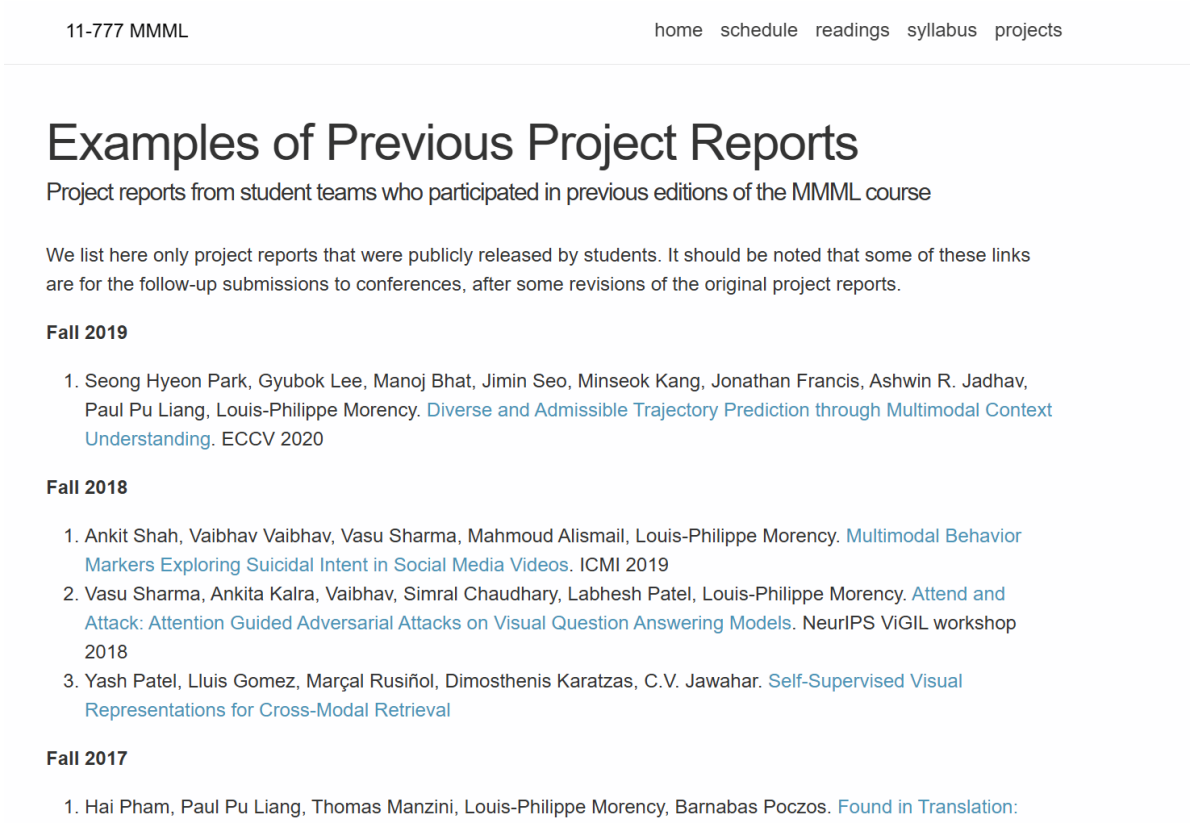
What will I learn with a recursive model?

Appendix: Some examples of Previous Projects

Project Examples

See the Fall 2023 course website:

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



11-777 MMML home schedule readings 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. Ankit Shah, Vaibhav Vaibhav, Vasu Sharma, Mahmoud Alismail, Louis-Philippe Morency. Multimodal Behavior Markers Exploring Suicidal Intent in Social Media Videos. ICMI 2019
2. 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
3. Yash Patel, Lluis Gomez, Marçal Rusiñol, Dimosthenis Karatzas, C.V. Jawahar. Self-Supervised Visual Representations for Cross-Modal Retrieval

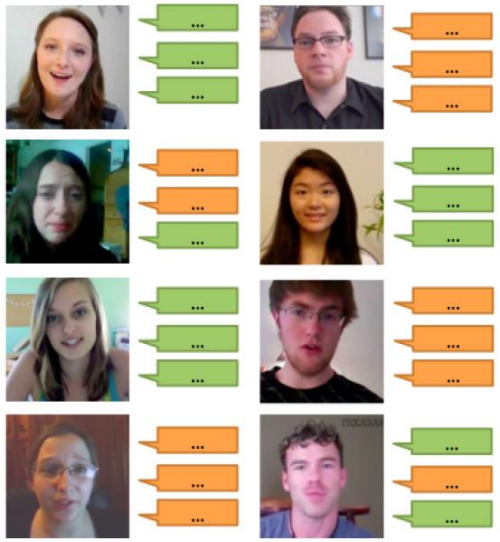
Fall 2017

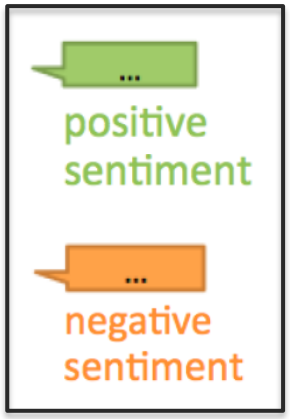
1. Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, Barnabas Poczos. Found in Translation:

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?

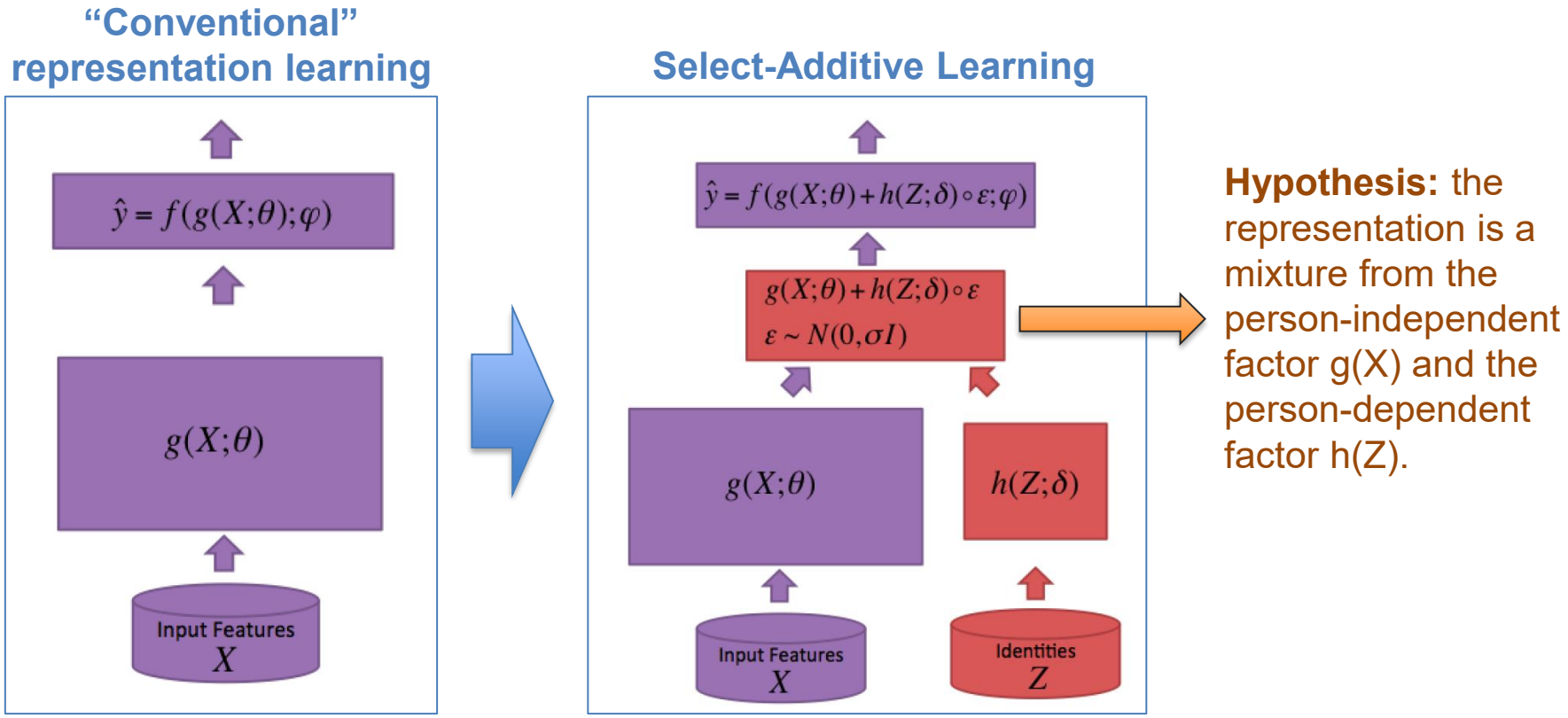
- ✔ *Smile* -> positive sentiment
- ✔ *Frown* -> negative sentiment
- ✔ *nod* -> positive sentiment
- ✗ *Wearing glasses* -> negative sentiment

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



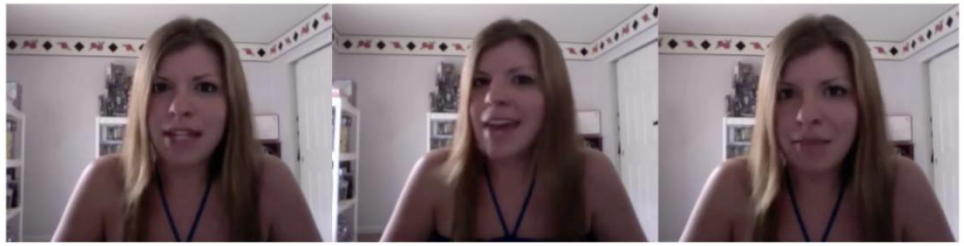
Hypothesis: the representation is a mixture from the person-independent factor g(X) and the person-dependent factor h(Z).

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

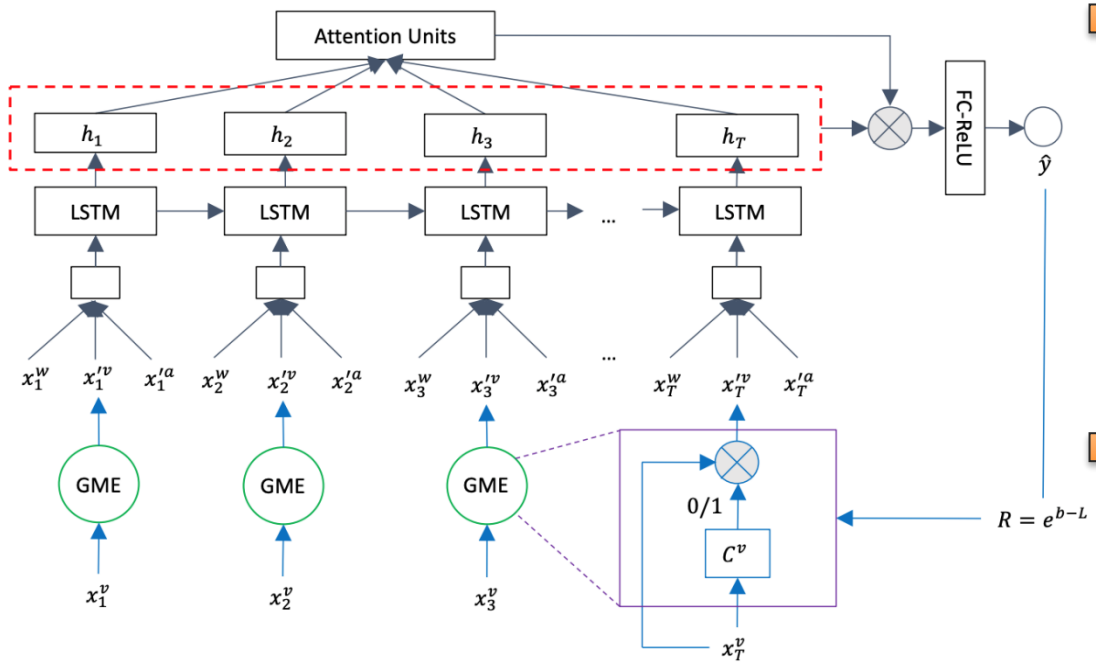
How can we build an interpretable model that estimates modality and temporal importance, and learns to attend to important information?

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

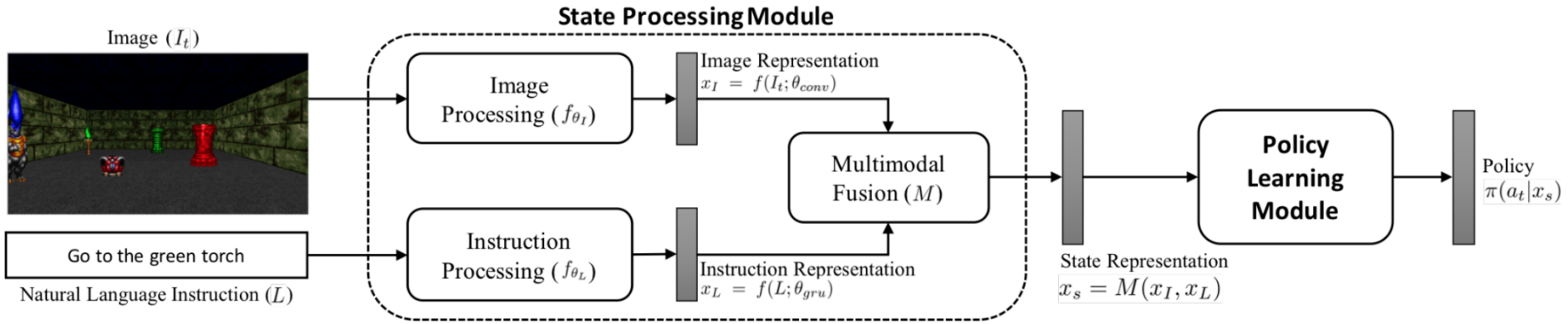
Project Example: Instruction Following

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

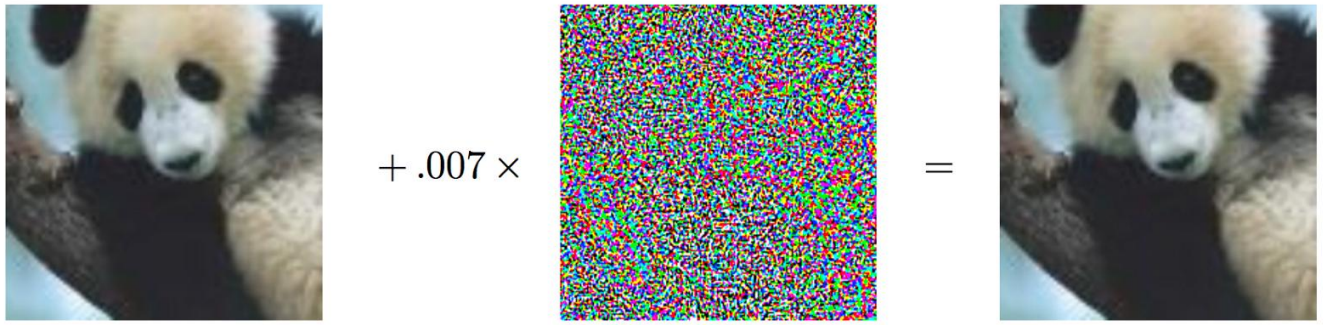


Hypothesis: Gated attention learns to ground and compose attributes in natural language with the image features. e.g. learning grounded representations for ‘green’ and ‘torch’.

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

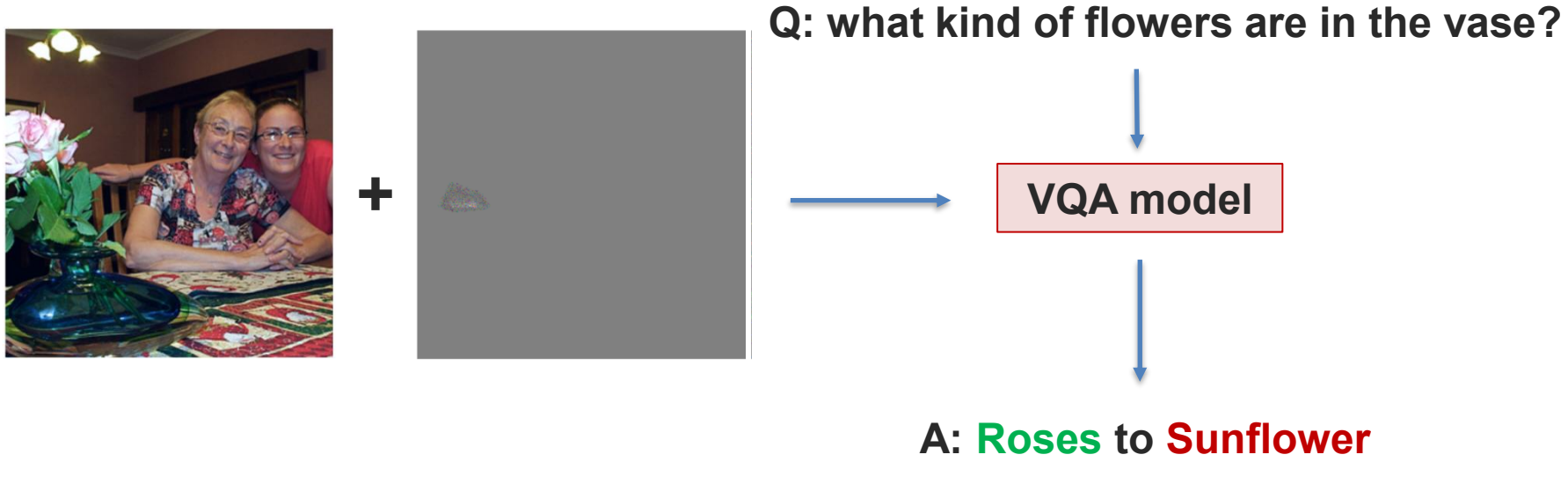
Research task: Adversarial Attacks on VQA models
Datasets: VQA
Main idea: Test the robustness of VQA models to adversarial attacks on the image.



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.

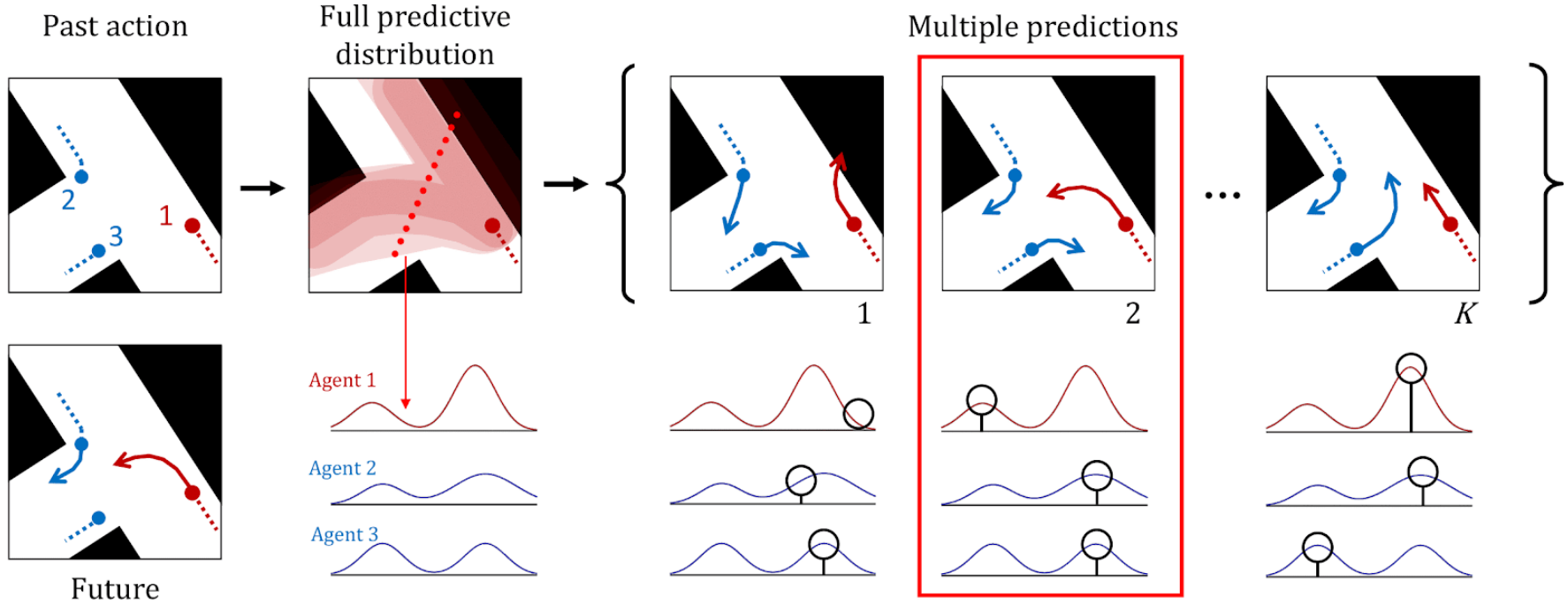


How can we design a targeted attack on images in VQA models, which will help in assessing robustness of existing models?

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

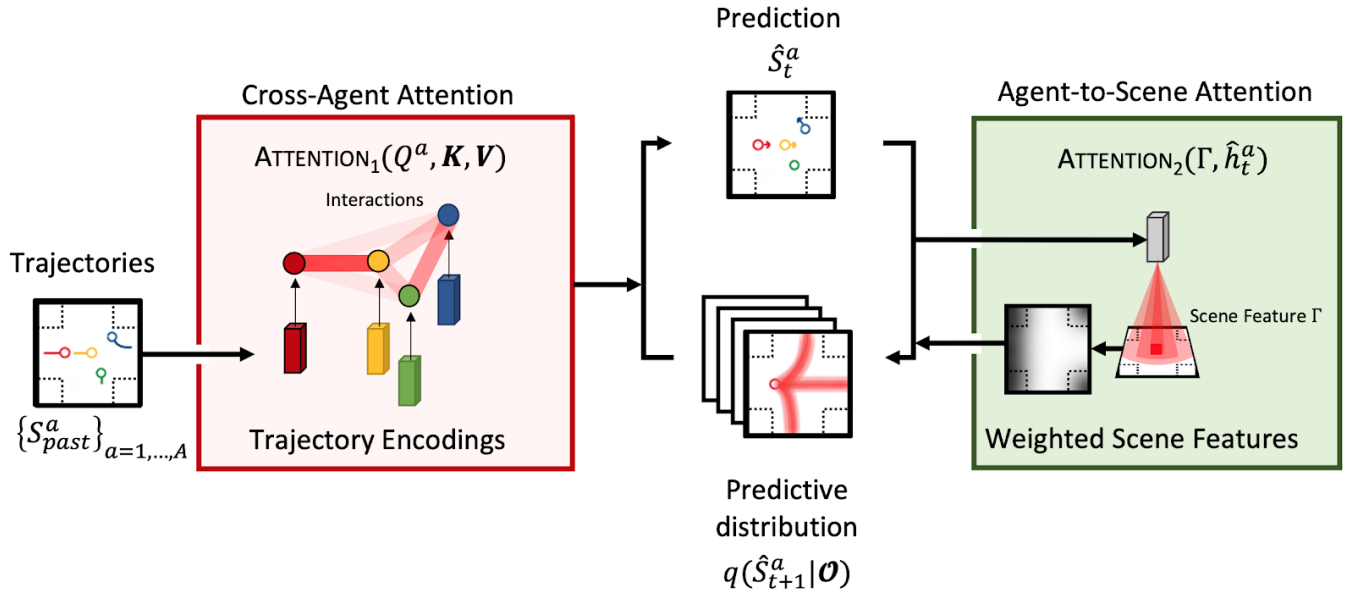
Research task: Multiagent trajectory forecasting for autonomous driving
Datasets: Argoverse and Nuscenes autonomous driving datasets
Main idea: Build a model that understands the environment and multiagent trajectories and predicts a set of multimodal future trajectories for each agent.



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