Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 10.2: Research Trends in Multimodal ML

Louis-Philippe Morency

Research Trends in Multimodal ML

- Abstraction and logic
- Multimodal reasoning
- Towards causal inference
- Understanding multimodal models
- Commonsense and coherence
- Social impact - fairness and misinformation
- Emotional and engaging interactions
- Multi-lingual multimodal grounding

Abstraction and Logic

Learning by Abstraction: The Neural State Machine

NEW paper

How to solve this question using visual reasoning?

What is the red fruit inside the bowl to the right of the coffee maker?

1. Given an **image**, generate a probabilistic **scene graph** that captures the semantic concepts.

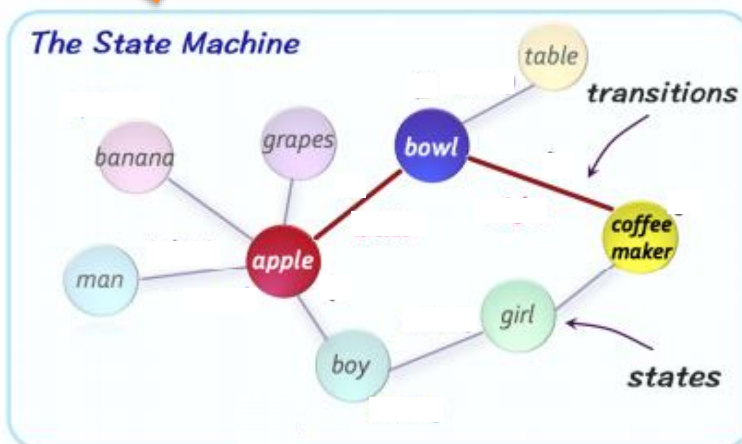
2. Treat the graph as a **state machine** and simulate iterative computation over it to *answer questions* or *draw inferences*.

3. Natural language questions are translated into *soft instructions* and used to perform sequential reasoning over the scene graph/state machine.

Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine.“ NeurIPS 2019

Learning by Abstraction: The Neural State Machine

Detect objects and create proximity graph

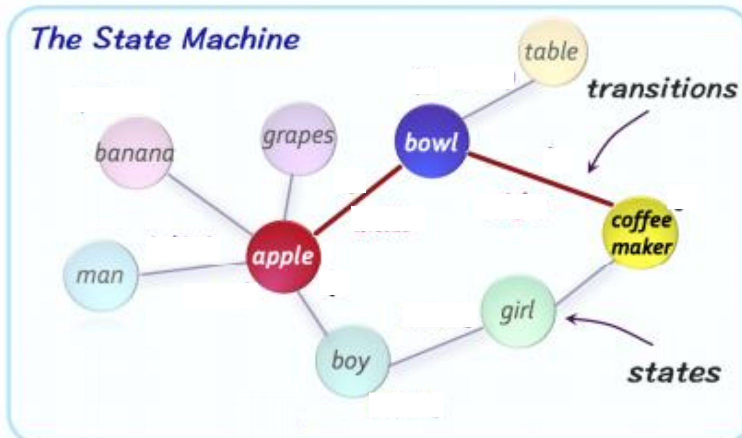


What is the red fruit inside the bowl to the right of the coffee maker?

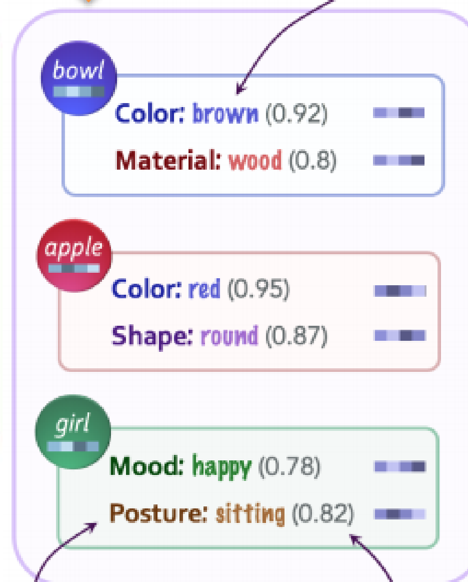
Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine.“ NeurIPS 2019

Learning by Abstraction: The Neural State Machine

Pre-trained an alphabet of concepts (Visual Genome)



alphabet (concepts)



What is the red fruit inside the bowl to the right of the coffee maker?

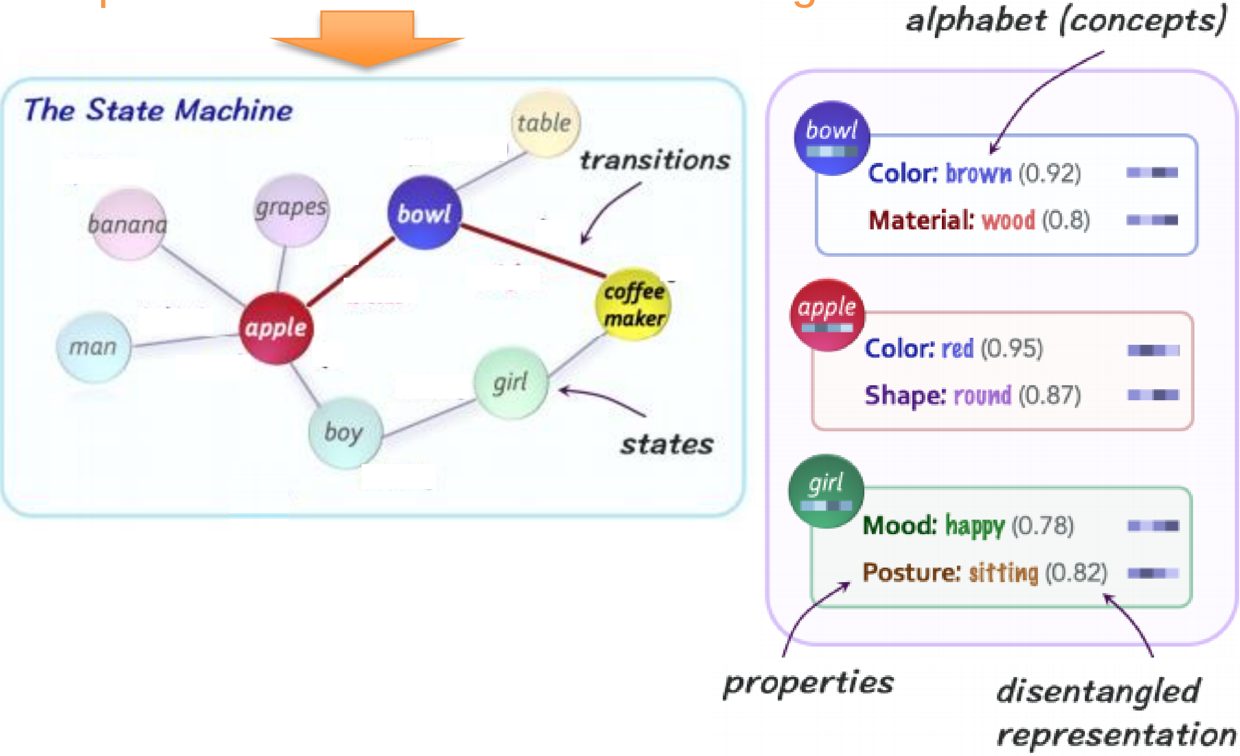
Manually grouped by "properties"

Probabilities computed at runtime for each object instance

Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine." NeurIPS 2019

Learning by Abstraction: The Neural State Machine

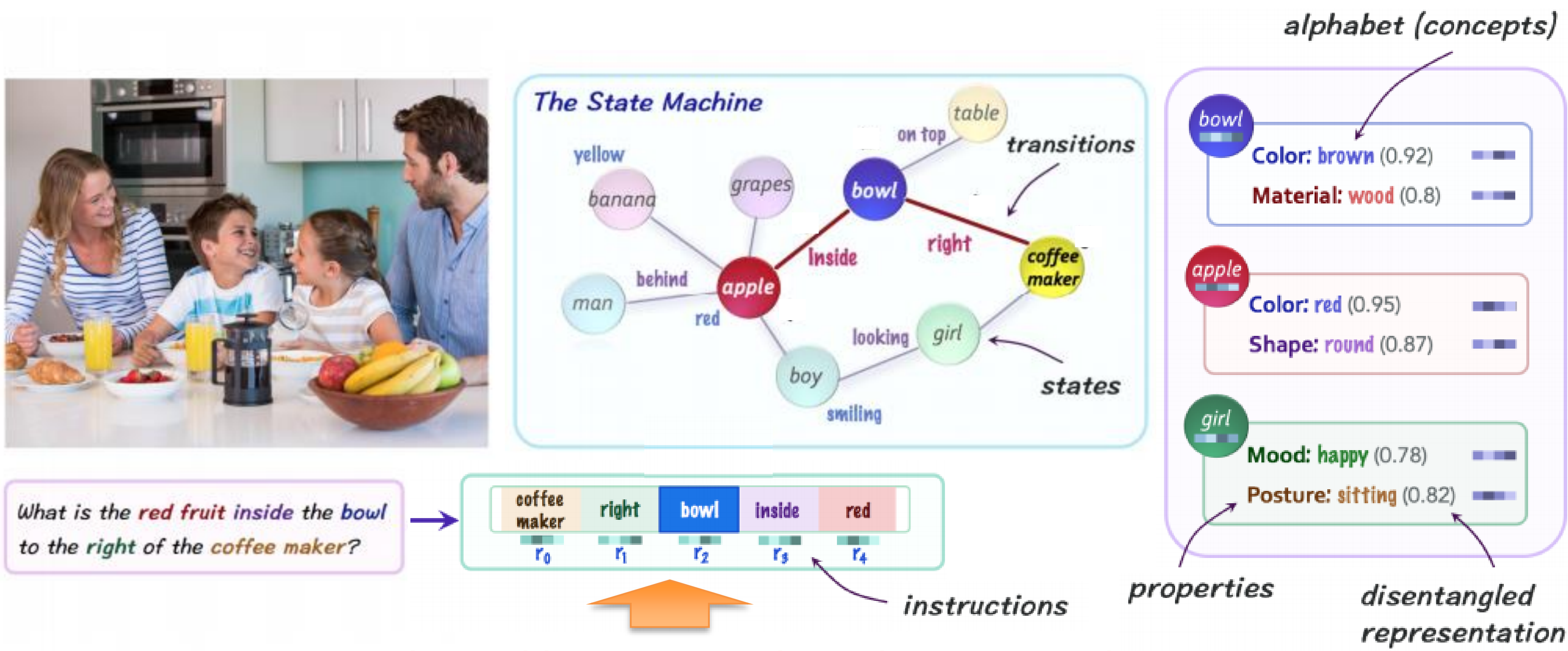
Predefined an alphabet of relations
and compute probabilities for each directed edges



What is the red fruit inside the bowl to the right of the coffee maker?

Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine." NeurIPS 2019

Learning by Abstraction: The Neural State Machine

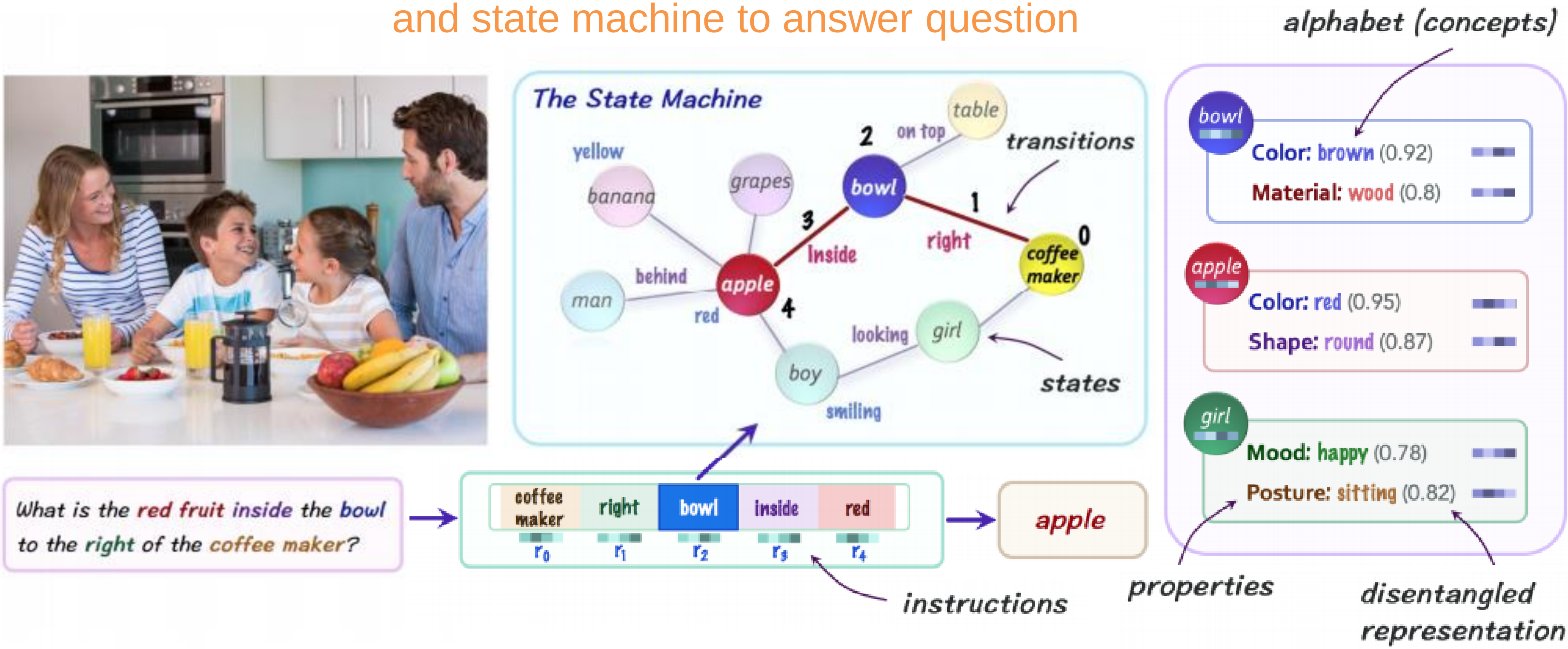


Translate each word in a concept-based representation
and group in a fixed number of instruction steps

Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine." NeurIPS 2019

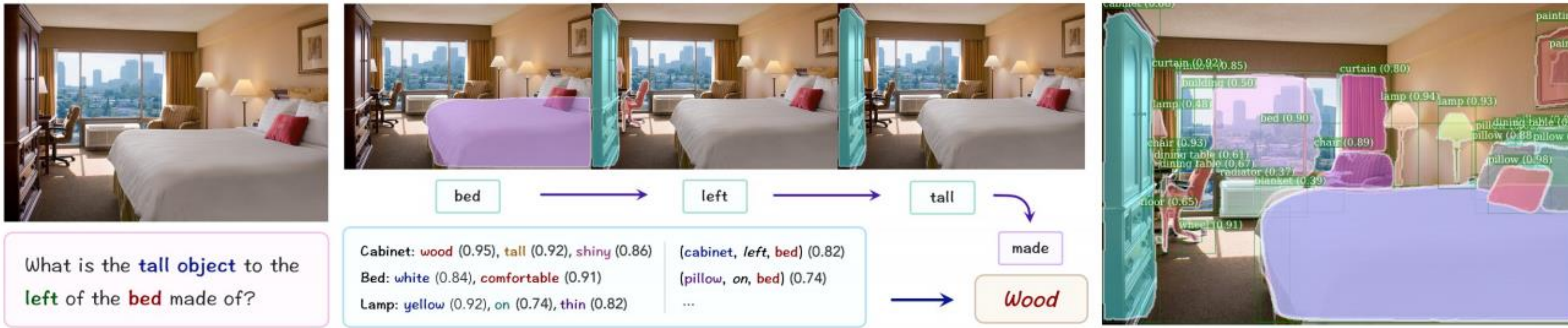
Learning by Abstraction: The Neural State Machine

Finally, perform reasoning using instructions
and state machine to answer question



Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine." NeurIPS 2019

Learning by Abstraction: The Neural State Machine



1. Compute the scene graph (blue boxes & image on the right)

2. Convert the question into a sequence of instructions (bed, left, tall, made)

3. Reason over the scene graph by attending to the relevant nodes using the instructions.

Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine." NeurIPS 2019

Learning by Abstraction: The Neural State Machine

Content Generalization

training	testing
Only questions that **do not** refer to any type of food or animal (do not include any word from these categories)	Only questions that refer to foods or animals (include a word from one of these categories)

Structure Generalization

training	testing
What is the <obj> covered by?	What is covering the <obj>?
Is there a <obj> in the image?	Do you see any <obj>s in the photo?
What is the <obj> made of?	What material makes up the <obj>?
What's the name of the <obj> that is <attr>?	What is the <attr> <obj> called?

Model	Content	Structure
Global Prior	8.51	14.64
Local Prior	12.14	18.21
Vision	17.51	18.68
Language	21.14	32.88
Lang+Vis	24.95	36.51
BottomUp [5]	29.72	41.83
MAC [40]	31.12	47.27
NSM	**40.24**	**55.72**

Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine.“ NeurIPS 2019

VQA under the Lens of Logic

Image

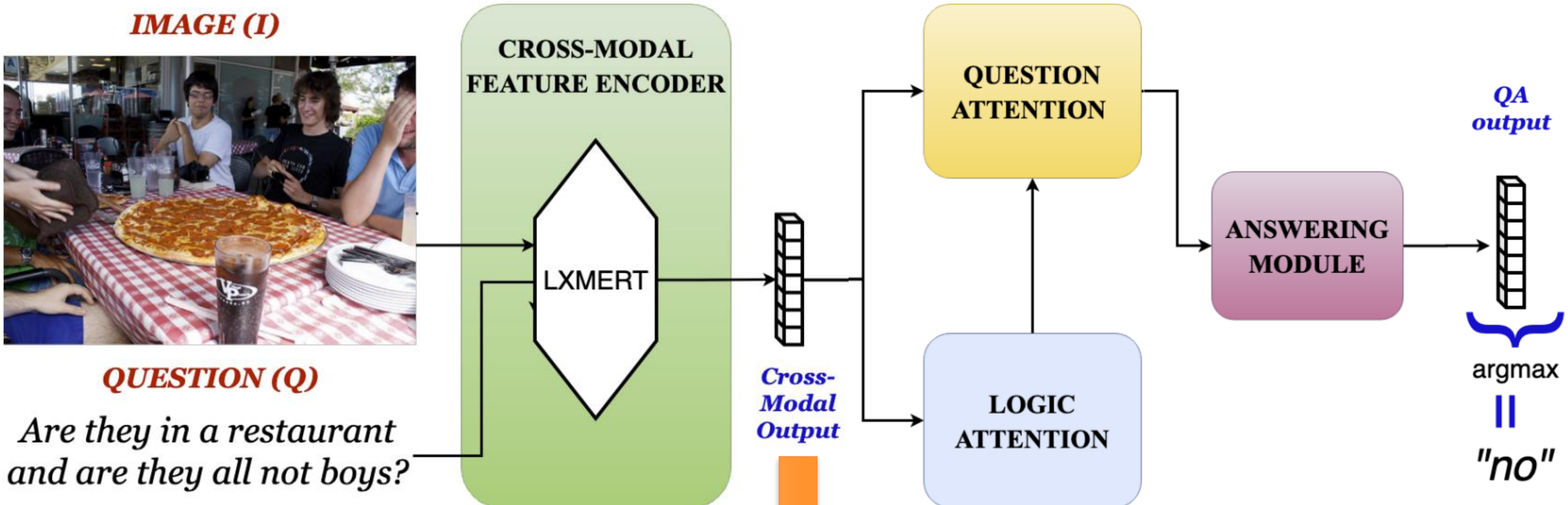
	Question	Predicted Answer by SOTA
VQA		
Q_1:	Is there beer?	YES (0.96)
Q_2:	Is the man wearing shoes?	NO (0.90)
VQA-Compose		
$\neg Q_2$:	Is the man *not* wearing shoes?	NO (0.80)
$\neg Q_2 \wedge Q_1$	Is the man *not* wearing shoes *and* is there beer?	NO (0.62)
$Q_1 \wedge C$	Is there beer and does this seem like a man bending over to look inside of a fridge?	NO (1.00)
VQA-Supplement		
$\neg Q_2 \vee B$	Is the man not wearing shoes or is there a clock?	NO (1.00)
$Q_1 \wedge anto(B)$	Is there beer and is there a wine glass?	YES (0.84)

New datasets

Gokhale, Tejas, et al. "VQA-LOL: Visual question answering under the lens of logic.“, ECCV 2020

VQA under the Lens of Logic

Attention over question types:
Yes-No, Number, or Other



1. Contextualized image embedding
2. Contextualized question embedding
3. Cross-model embedding

Attention over logic connectives:
AND, OR, NOT, no connective

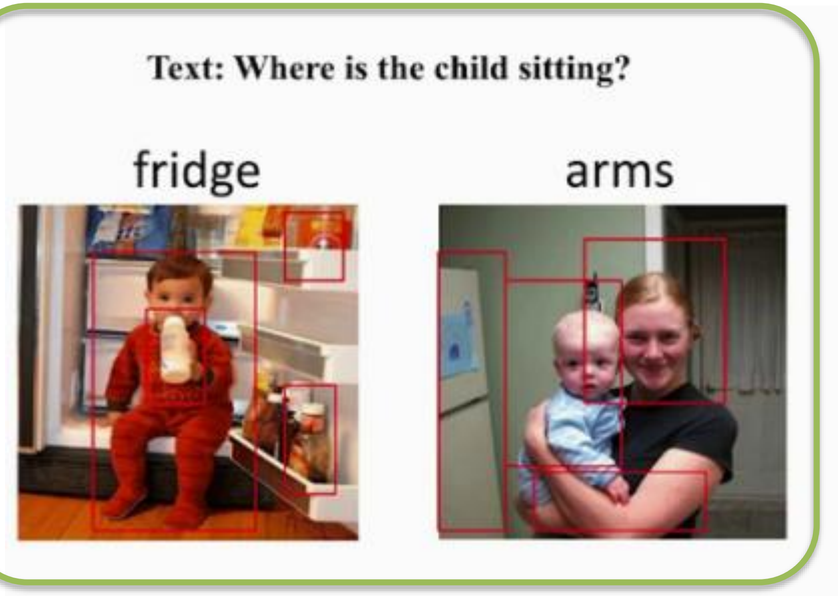
Gokhale, Tejas, et al. "VQA-LOL: Visual question answering under the lens of logic.", ECCV 2020

Multimodal Reasoning

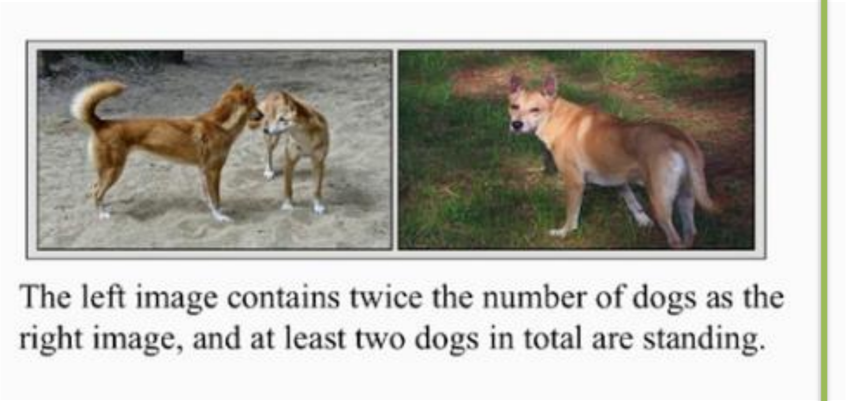
Cross-Modality Relevance for Reasoning on Language and Vision

NEW paper

Visual Question Answering



Natural Language for Visual Reasoning



Solving these problems requires:

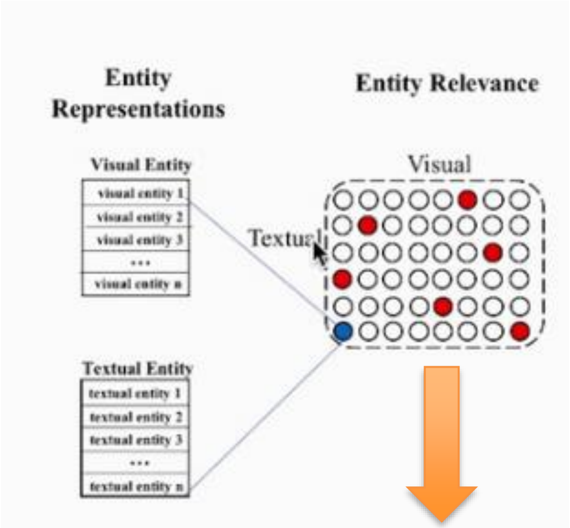
(1) Knowing relevance (aka, alignment) between visual and language entities

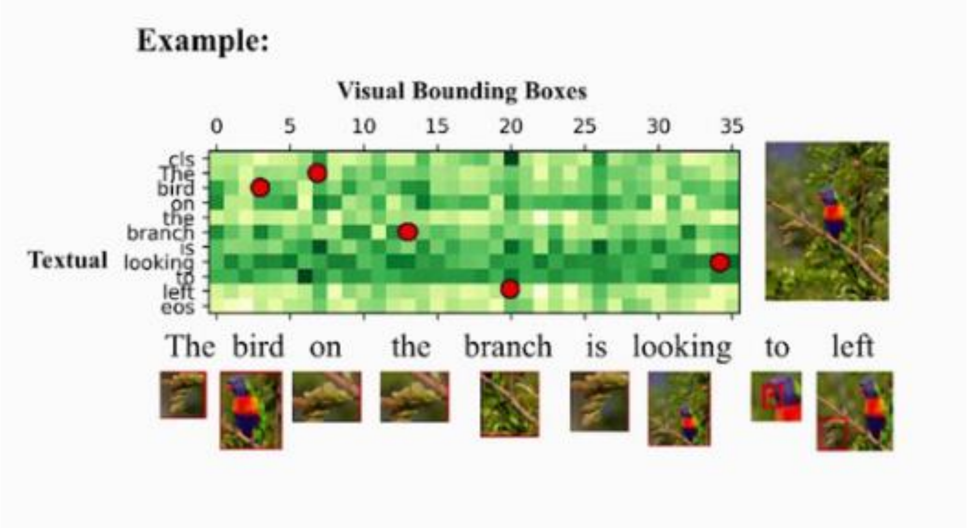
(2) Knowing relevance between visual pairs and language pairs

Cross-Modality Relevance for Reasoning on Language and Vision, ACL 2020

Cross-Modality Relevance for Reasoning on Language and Vision

Computing **Cross Modality Relevance** affinity matrix



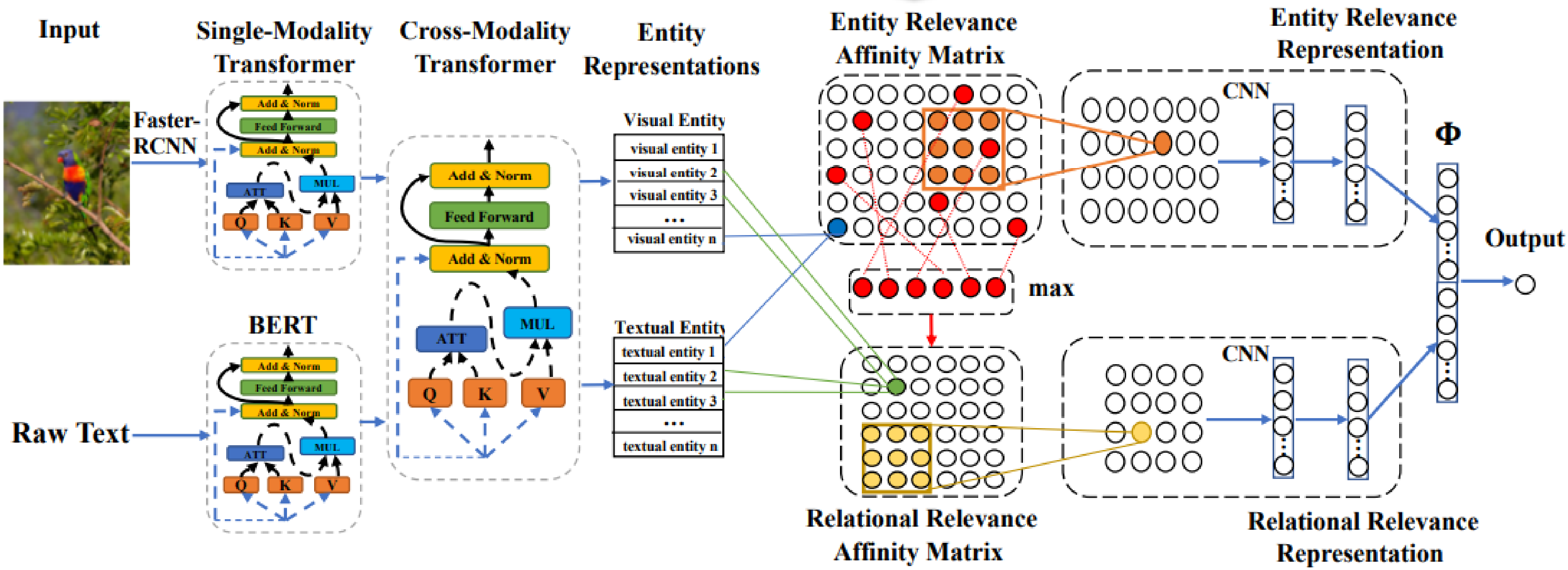


Similar bilinear models

Cross-Modality Relevance for Reasoning on Language and Vision, ACL 2020

Cross-Modality Relevance for Reasoning on Language and Vision

Alignment between visual and language entities



Alignment between visual and language pairs

Cross-Modality Relevance for Reasoning on Language and Vision, ACL 2020

Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog

NEW-ish paper

Hypothesis: The failure of visual dialog is caused by the inherent weakness of single-step reasoning.

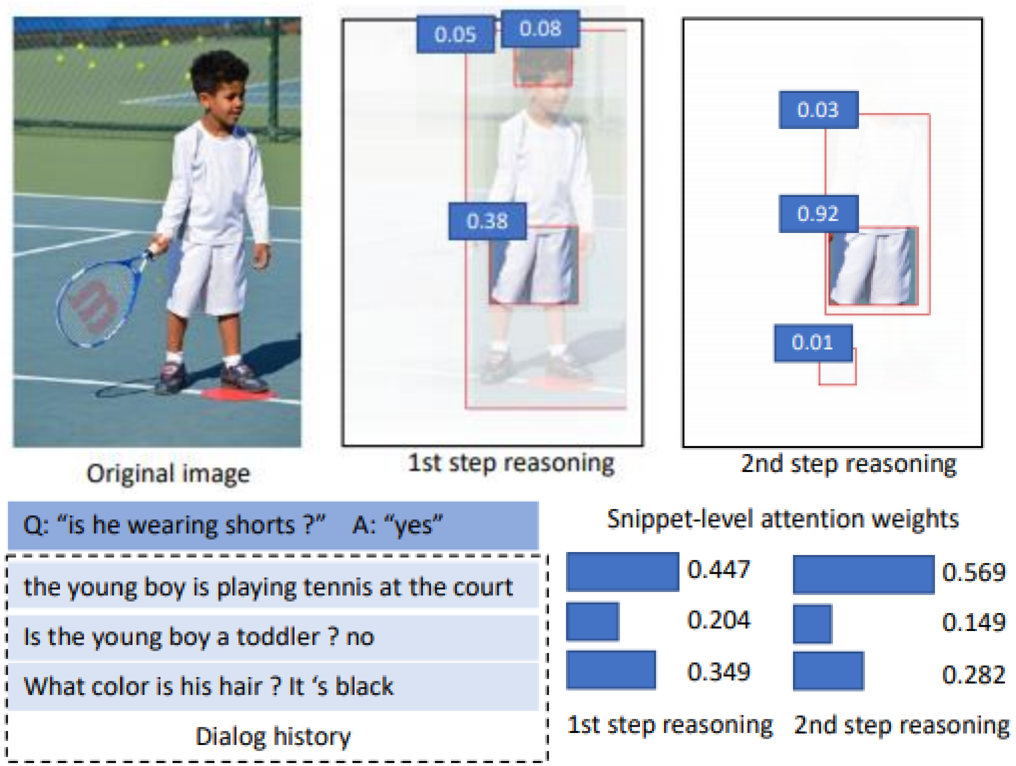
Intuition: Humans take a first glimpse of an image and a dialog history, before *revisiting* specific parts of the image/text to understand the multimodal context.

Proposal: Apply *Multi-step reasoning* to visual dialog by using a recurrent (aka multi-step) version of attention (aka reasoning). This is done on both text and questions (aka, dual).

➡ Recurrent Dual Attention Network

Gan, Zhe, et al. "Multi-step reasoning via recurrent dual attention for visual dialog." ACL 2019

Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog



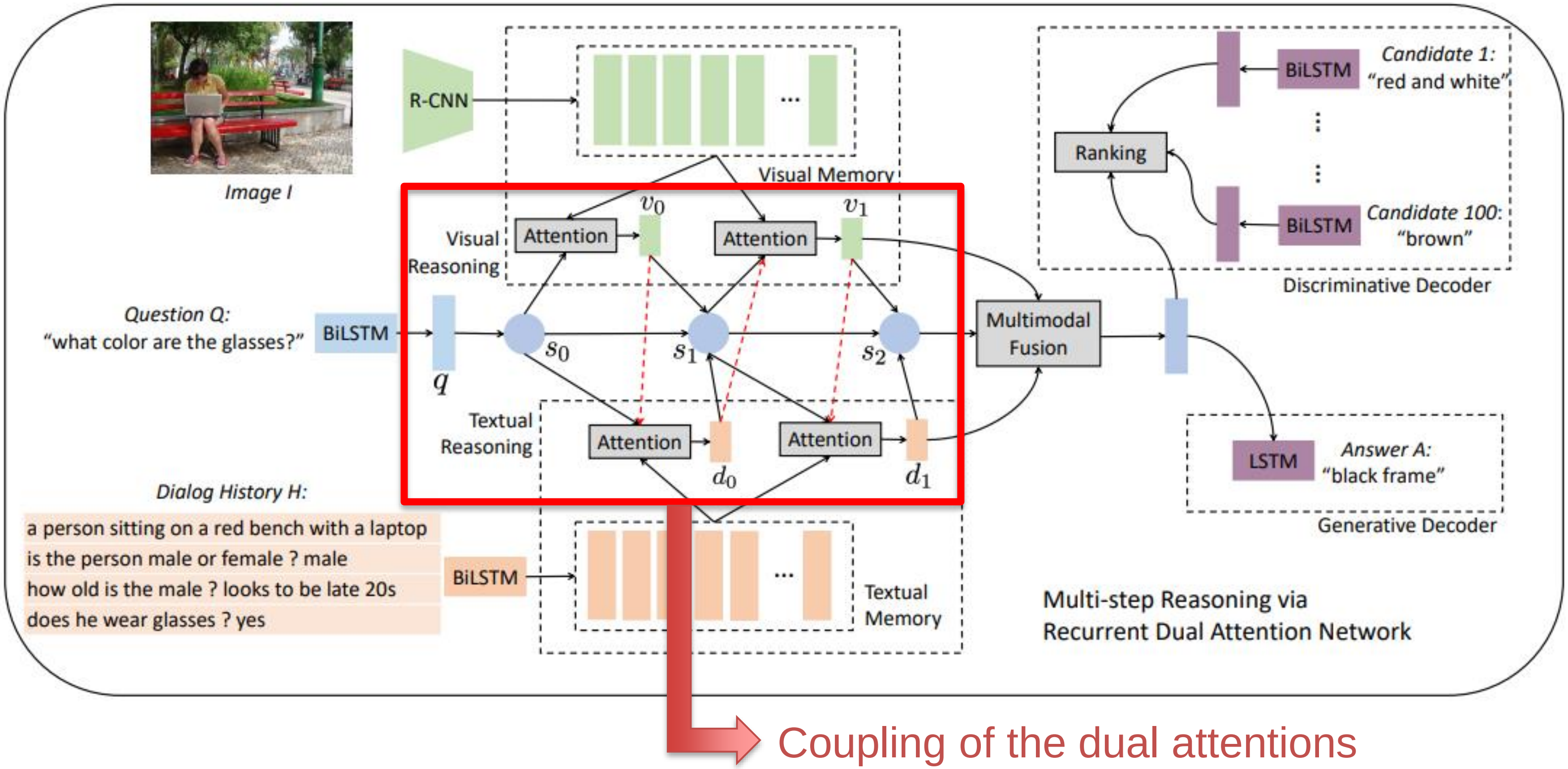
1st Step Reasoning: Attend to *all relevant* objects and dialog turns.

2nd Step Reasoning: Narrow down to context relevant regions (shorts, young boy).

In the 2nd step, the attention becomes sharper.

Gan, Zhe, et al. "Multi-step reasoning via recurrent dual attention for visual dialog." ACL 2019

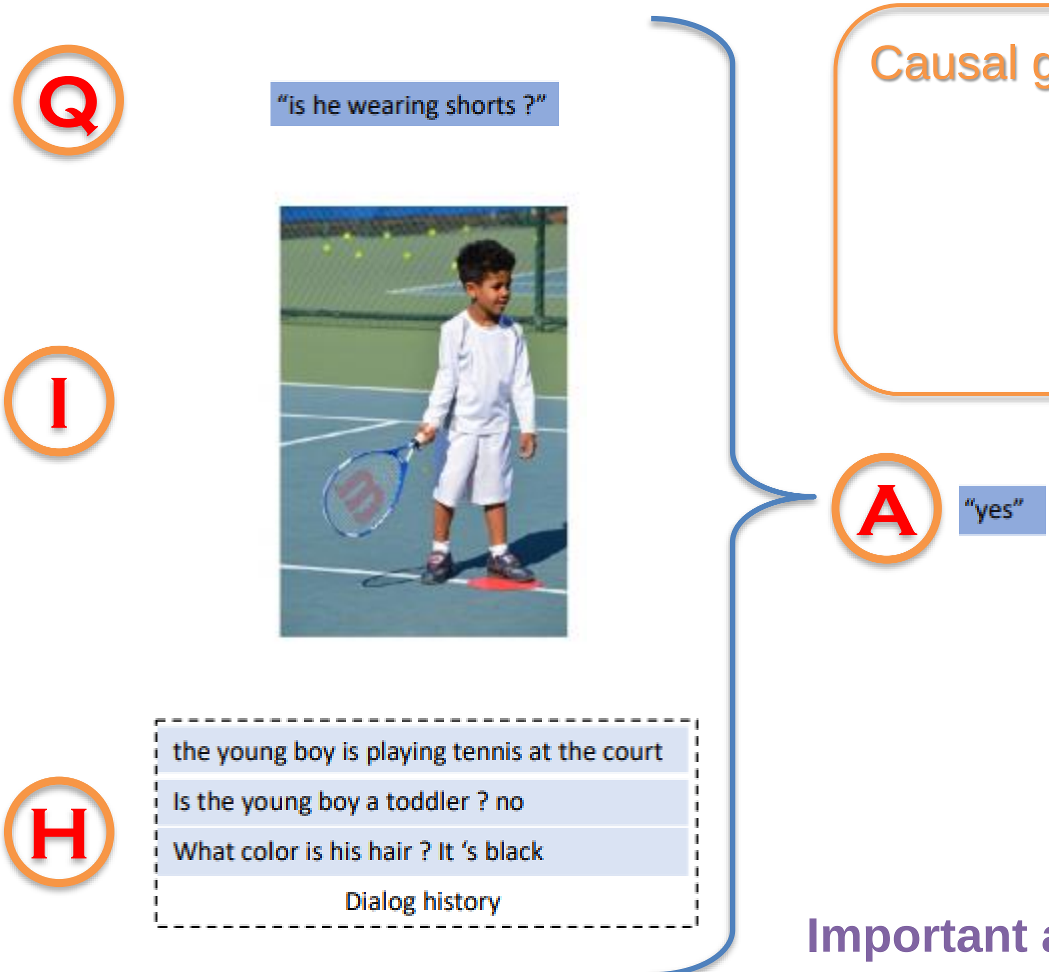
Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog



Gan, Zhe, et al. "Multi-step reasoning via recurrent dual attention for visual dialog." ACL 2019

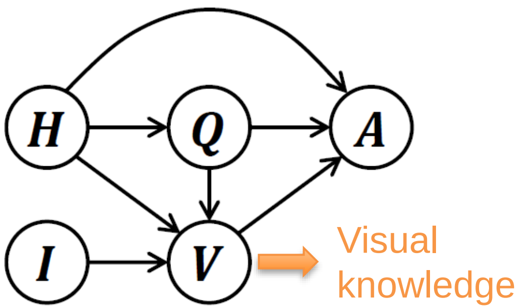
Towards Causal Inference

Visual Dialogue Expressed with Causal Graph



Causal graph: acyclic graph where nodes denote variables and edges denote causal relationships

How to represent this visual dialogue problem?



Important assumption: the output of a neural network is the *effect* of the input (the *cause*)

Two Causal Principles for Improving Visual Dialog

NEW paper

This paper identifies two causal principles that are holding back VisDial models.

1. **Harmful shortcut bias** between dialog history (H) and the answer (A)

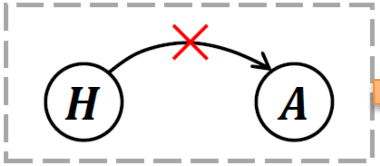
2. **Unobserved confounder** between H, Q and A leading to spurious correlations.

By identifying and addressing these principles in a model-agnostic manner, they are able to promote any VisDial model to SOTA levels.

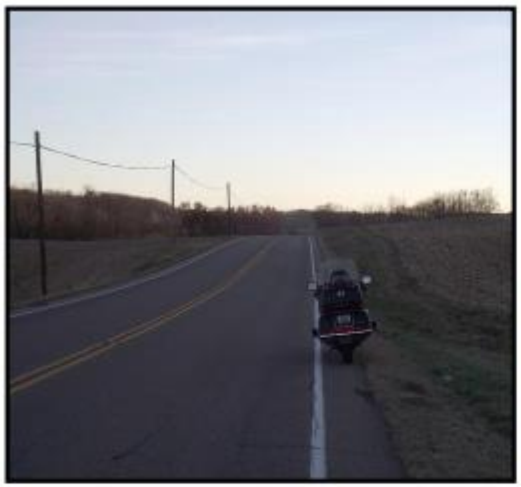
Qi, Jiaxin, et al. "Two causal principles for improving visual dialog.“ CVPR 2020

Two Causal Principles for Improving Visual Dialog

Principle 1: Harmful shortcut bias between dialog history (H) and the answer (A)



Dataset bias example:



H

H_0:A motorcycle parked on the road site

Q_1:Is the photo in color? A_1:It is in color

Q_2:Is there any people? A_2:I don't see any people

Q_3: Any other motorcycles? A_3:No other motorcycles

Q_4: Is it night? A_4:It is either morning or near sunset

Q_5: What color of motorcycles? A_5:Dark colored

Q_6:Is there trees? A_6:There are trees, in the background

Q_7:Any other vehicles?

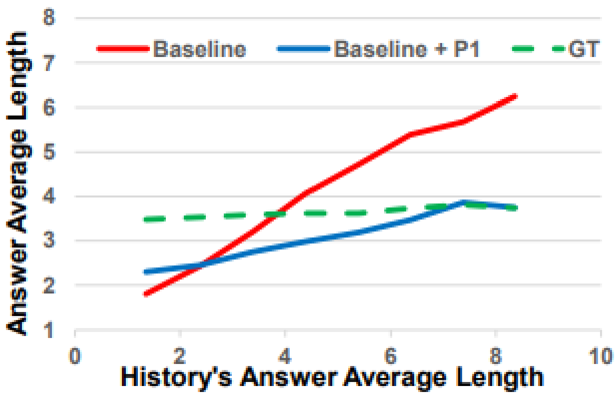
GT Answer: No other vehicles

Ranked A (Baseline)

1.No other vehicles
2.There are no animals
3.I don't see any other building
⋮

Ranked A (Baseline + P1)

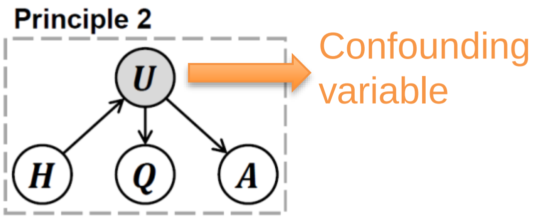
1.No
2.No other vehicles
3.Nope
⋮



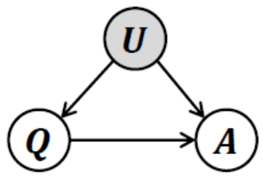
Qi, Jiaxin, et al. "Two causal principles for improving visual dialog.“ CVPR 2020

Two Causal Principles for Improving Visual Dialog

Principle 2: Unobserved confounder between H and A (as well as between H and Q) leading to spurious correlations.



Explaining confounding variable:



We may think that Q is primarily causing A, but U is a common cause for both Q and A

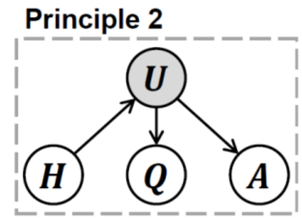
U has a *spurious* relation with Q and A

In our case, U is *unobserved*, and most likely because answerers (aka "users") could see the history.

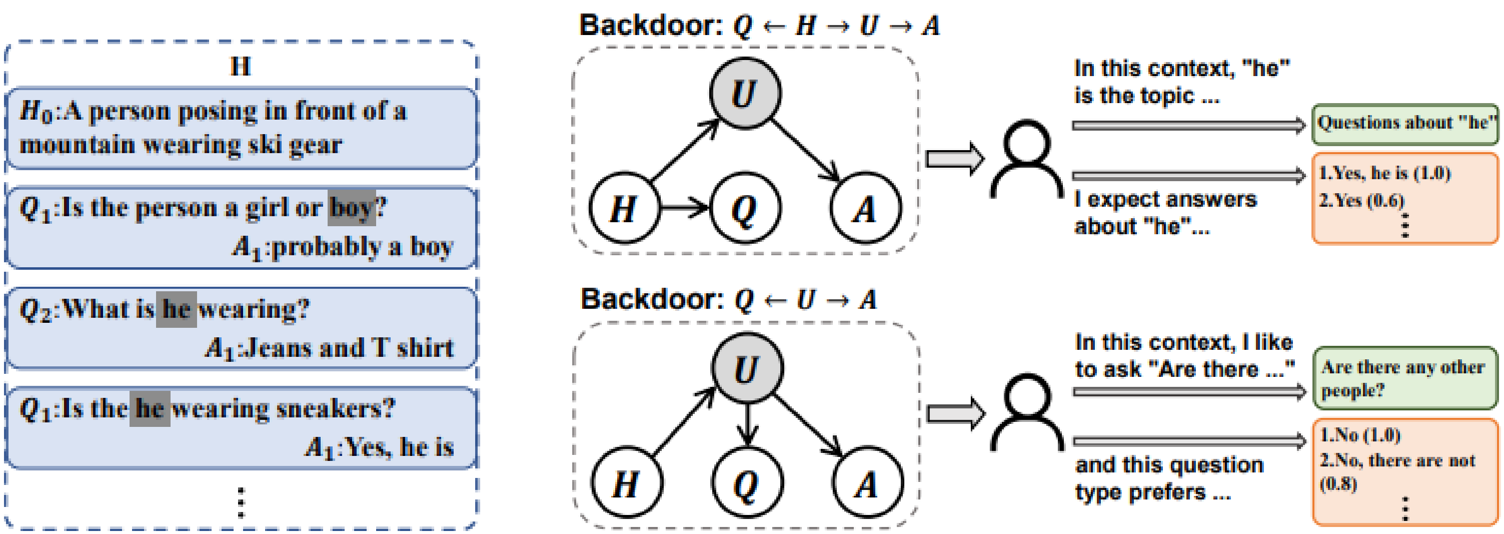
Qi, Jiaxin, et al. "Two causal principles for improving visual dialog." CVPR 2020

Two Causal Principles for Improving Visual Dialog

Principle 2: Unobserved confounder between H, Q and A leading to spurious correlations.



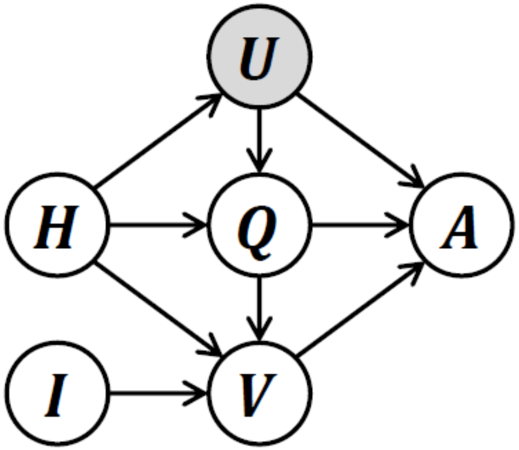
Dataset bias example:



Qi, Jiaxin, et al. "Two causal principles for improving visual dialog." CVPR 2020

Two Causal Principles for Improving Visual Dialog

Proposed method



1. Removes the **Harmful shortcut bias** between dialog history (H) and the answer (A)

2. Explicitly model the **unobserved confounder** between H, Q and A

Qi, Jiaxin, et al. "Two causal principles for improving visual dialog.“ CVPR 2020

Studying Biases in VQA Models

	Prediction
What is in the basket?	banana
What is contained in the basket?	pizza
What can be seen inside the basket?	remote
What does the basket mainly contain?	paper

Why one question was correctly answered and not the others?

VQA models may be finding spurious correlations (e.g., confounding variables)

Research idea: Try to remove visual objects to see if they are confounding variables. + Propose a new evaluation metric to measure it.

Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."

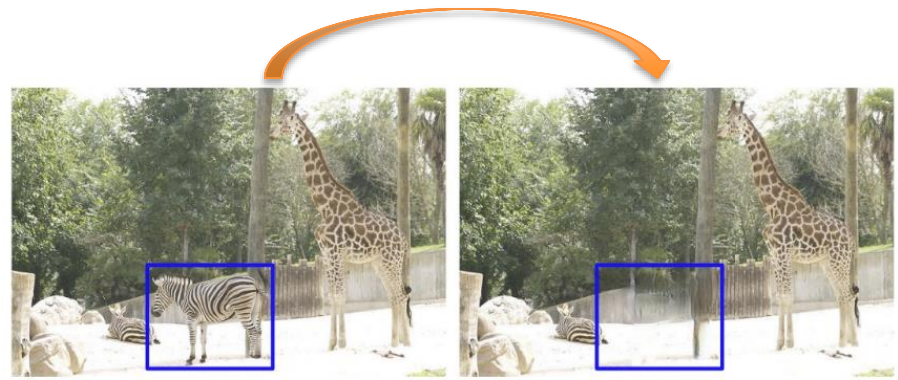
Studying Biases in VQA Models

Consistency **metric:** Study the change in performance when individual objects are removed from the image

➡ using GAN to manipulate the images

Q: Is this a kitchen?

A: no *toilet removed*; A: no



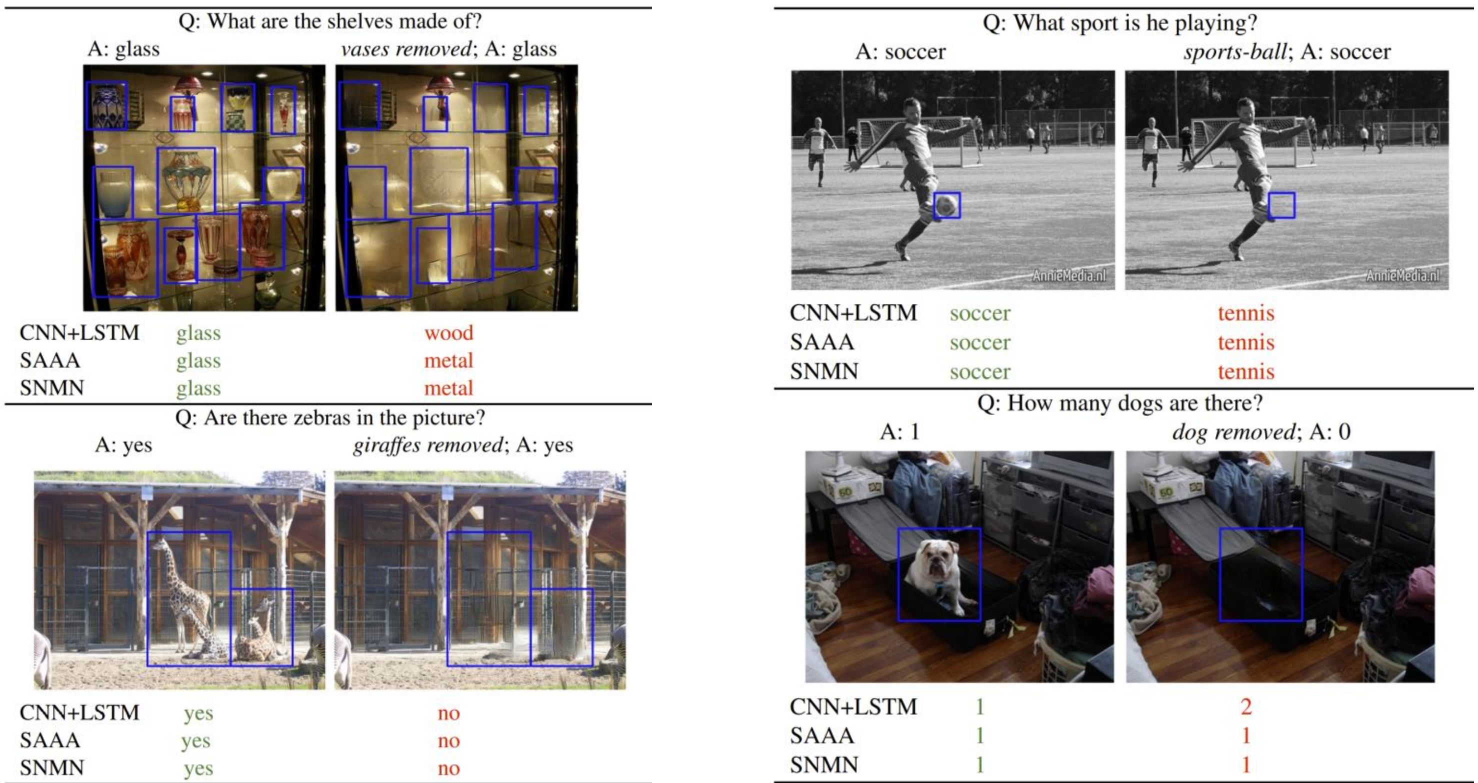
Q: How many zebras are there in the picture?

A: 2 *zebra removed* A: 1

Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."

Studying Biases in VQA Models

State-of-the-art models often exploit spurious correlations…



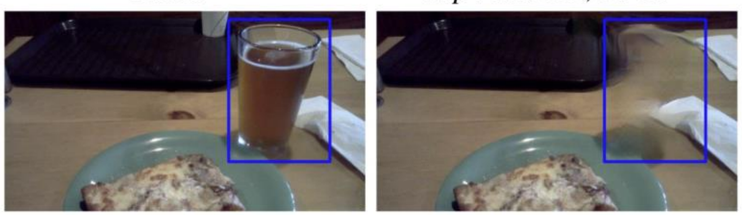
Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."

Studying Biases in VQA Models

Proposed solution: training the model on original VQA datasets plus synthetic datasets, consisting of images with removed objects.

Q:Is there a bowl on the table?

A: no — *cup removed*; A: no



	real	real+edit	real	real+edit
CL	no	no	yes	no
SAAA	no	no	yes	no
SNMN	no	no	yes	no

Q: How many people are in the water?

A: 1 — *person removed*; A: 0

	real	real+edit	real	real+edit
CL	1	1	1	0
SAAA	1	1	1	0
SNMN	1	1	1	0

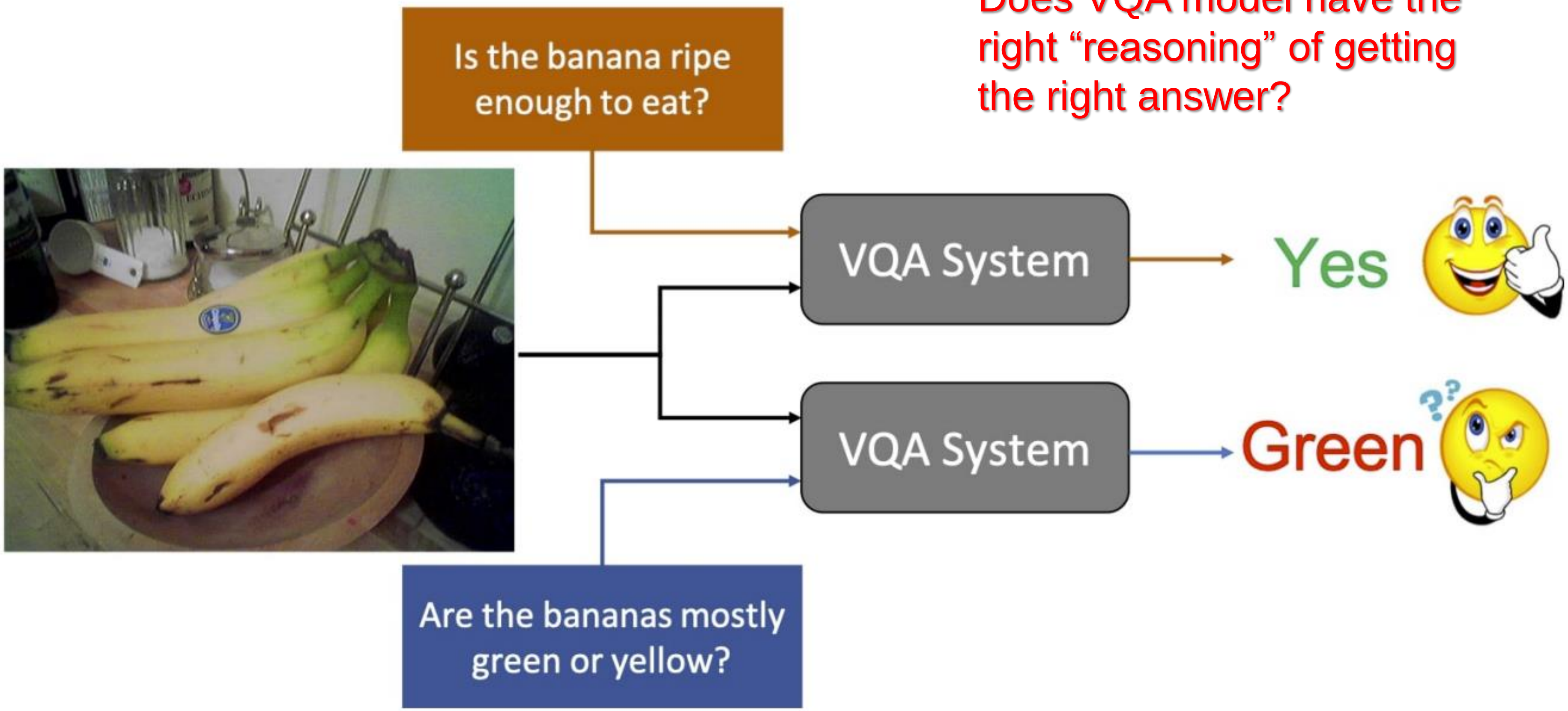
Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."

Understanding Multimodal Models

Introspecting VQA Models with Sub-Questions

NEW paper

Does VQA model have the right “reasoning” of getting the right answer?



Selvaraju, Ramprasaath R., et al. "SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions.", CVPR 2020

New Dataset

1. Select only the *Reasoning* questions (requires composition of perceptions and knowledge) from the VQA dataset

2. Add many *Perception* questions (recognize existence of visual objects) as sub-questions, to further validate VQA models.

Main Reasoning Question:
- Is this a keepsake photo? “Yes”

Perception Sub-questions:
- Is this a black and white photo? “Yes”
- Is the woman wearing a white veil and holding flowers? “Yes”
- Is the woman wearing a veil? “Yes”
- What is the woman next to the man wearing? “Gown”

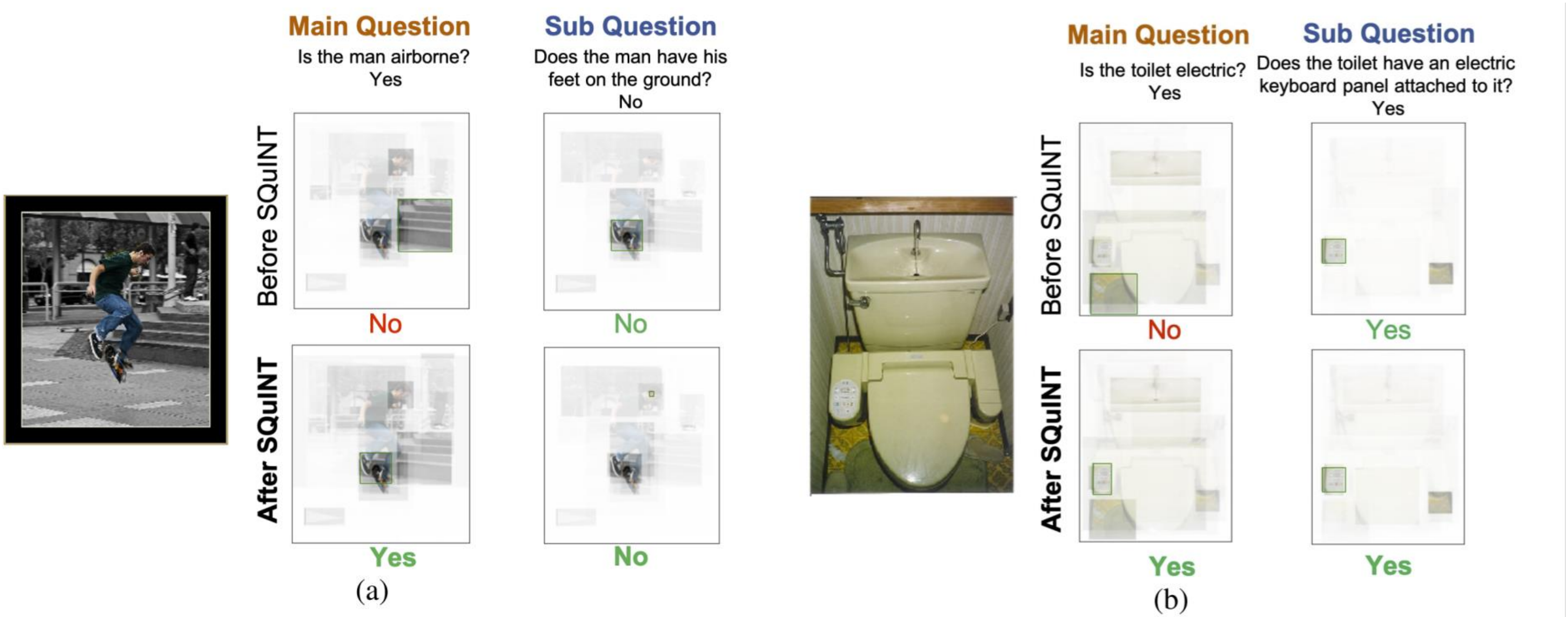
Main Reasoning Question:
- Is this giraffe at the zoo? “Yes”

Perception Sub-questions:
- Is the giraffe fenced in? “Yes”
- Is the grass shorter than 3 inches? “Yes”
- Is there a fence? “Yes”
- Is a fence around the giraffe? “Yes”

Selvaraju, Ramprasaath R., et al. "SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions.”, CVPR 2020

SQuINTing Model

Proposed method: Attend to the same region when answering both main questions and sub-questions.

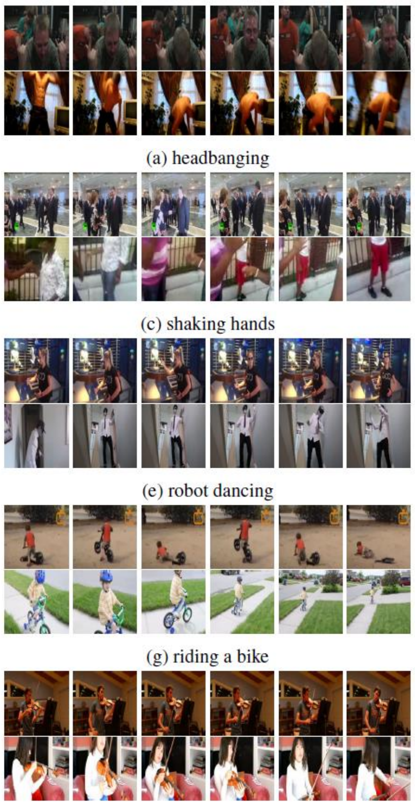


Selvaraju, Ramprasaath R., et al. "SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions.", CVPR 2020

What Makes Training Multi-modal Classification Networks Hard?

NEW paper

Kinetics dataset



Adding more modalities should always help?

Modalities:
RGB (video clips)
A (Audio features)
OF (optical flow - motion)

Dataset	Multi-modal	V@1	Best Uni	V@1	Drop
Kinetics	A + RGB	71.4	RGB	**72.6**	-1.2
	RGB + OF	71.3	RGB	**72.6**	-1.3
	A + OF	58.3	OF	**62.1**	-3.8
	A + RGB + OF	70.0	RGB	**72.6**	-2.6

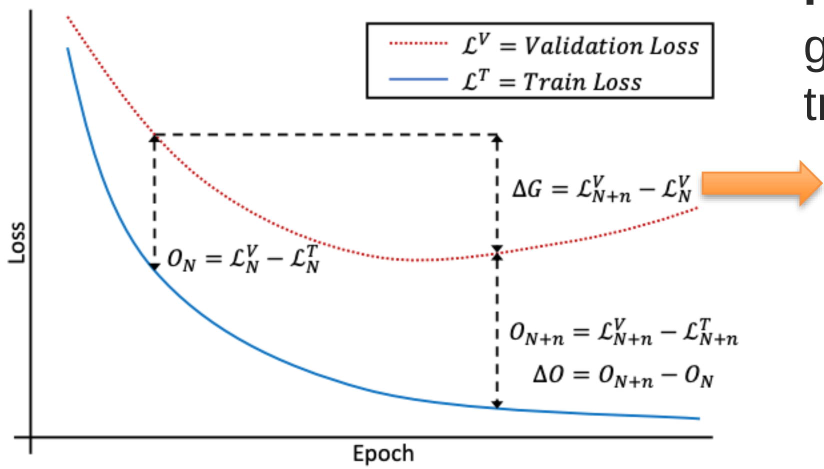
But sometimes multimodal doesn't help! **Why?**

Wang et al., What Makes Training Multi-modal Classification Networks Hard?. CVPR 2020

Training Multimodal Networks

2 possible explanations for drop in performance:

1. Multimodal networks are more prone to overfitting due to increased complexity
2. Different modalities overfit and generalize at different rates so training them jointly with a single optimization strategy may be sub-optimal



Key idea 1: compute overfitting-to-generalization ratio (OGR) between training checkpoints

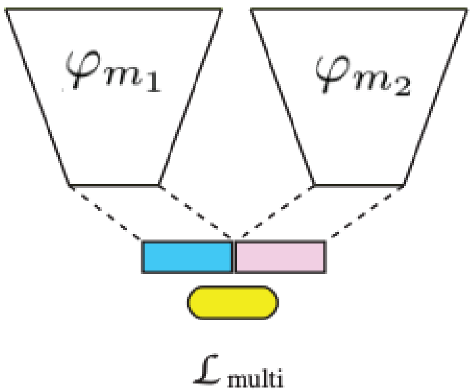
Gap between training and valid loss

OGR wrt each modality tells us how much to train that modality

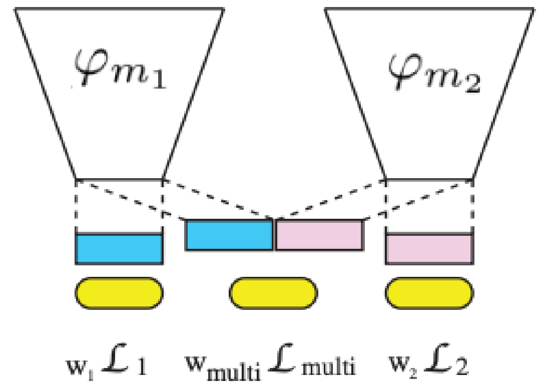
Wang et al., What Makes Training Multi-modal Classification Networks Hard?. CVPR 2020

Training Multimodal Networks

Conventional approach
(with late fusion)



Proposed approach



Key idea 2: Simultaneously train unimodal networks to estimate OGR wrt each modality

+ Reweight multimodal loss using unimodal OGR values

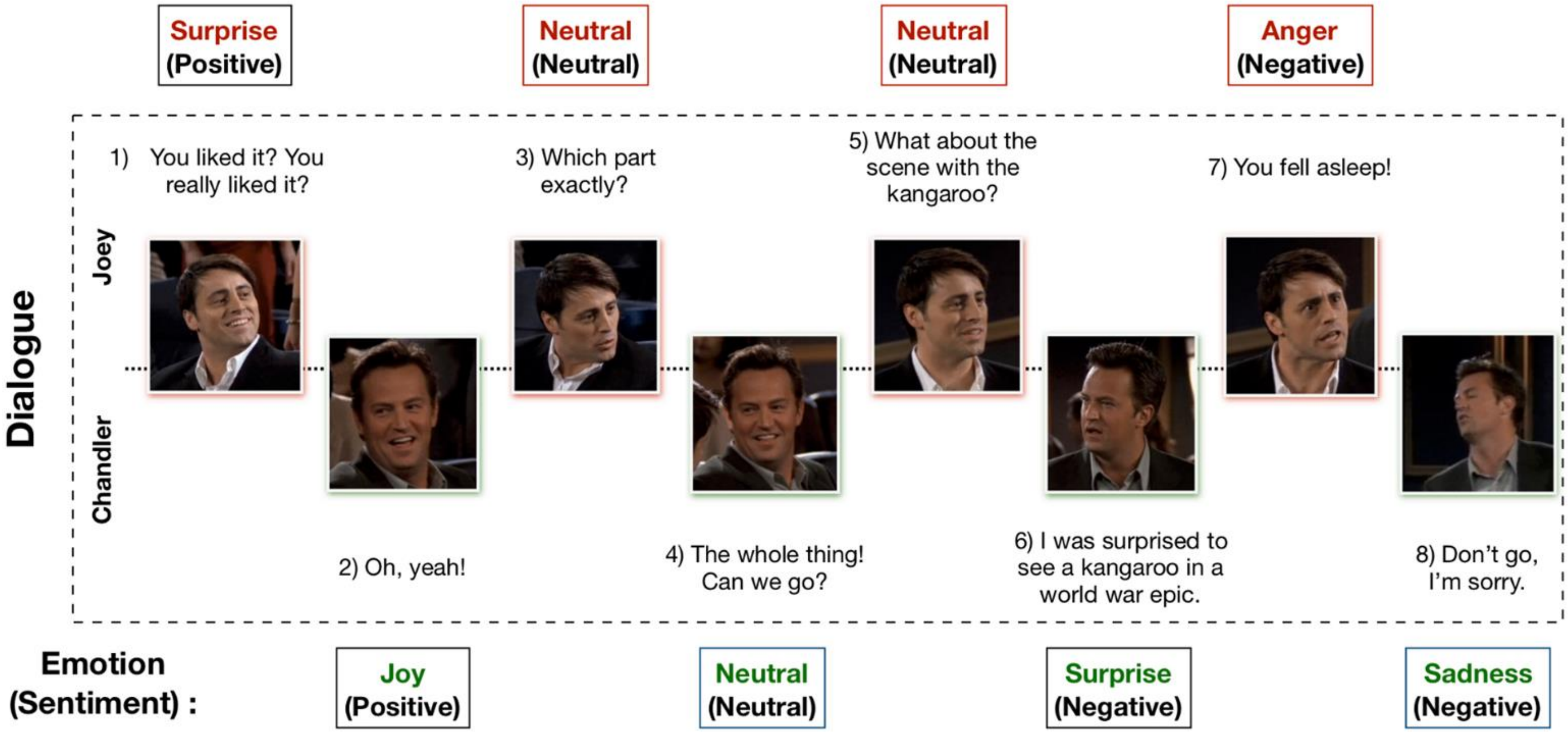
→ Allows to better balance generalization & overfitting rate of different modalities

Wang et al., What Makes Training Multi-modal Classification Networks Hard?. CVPR 2020

Commonsense and Coherence

NEW paper

Emotions are Often Context Dependent



"COSMIC: COmmonSense knowledge for eMotion Identification in Conversations", Findings of EMNLP 2020

Commonsense and Emotion Recognition

Proposed approach (COSMIC):

For each utterance, try to infer

- speaker's intention
- effect on the speaker/listener
- reaction of the speaker/listener

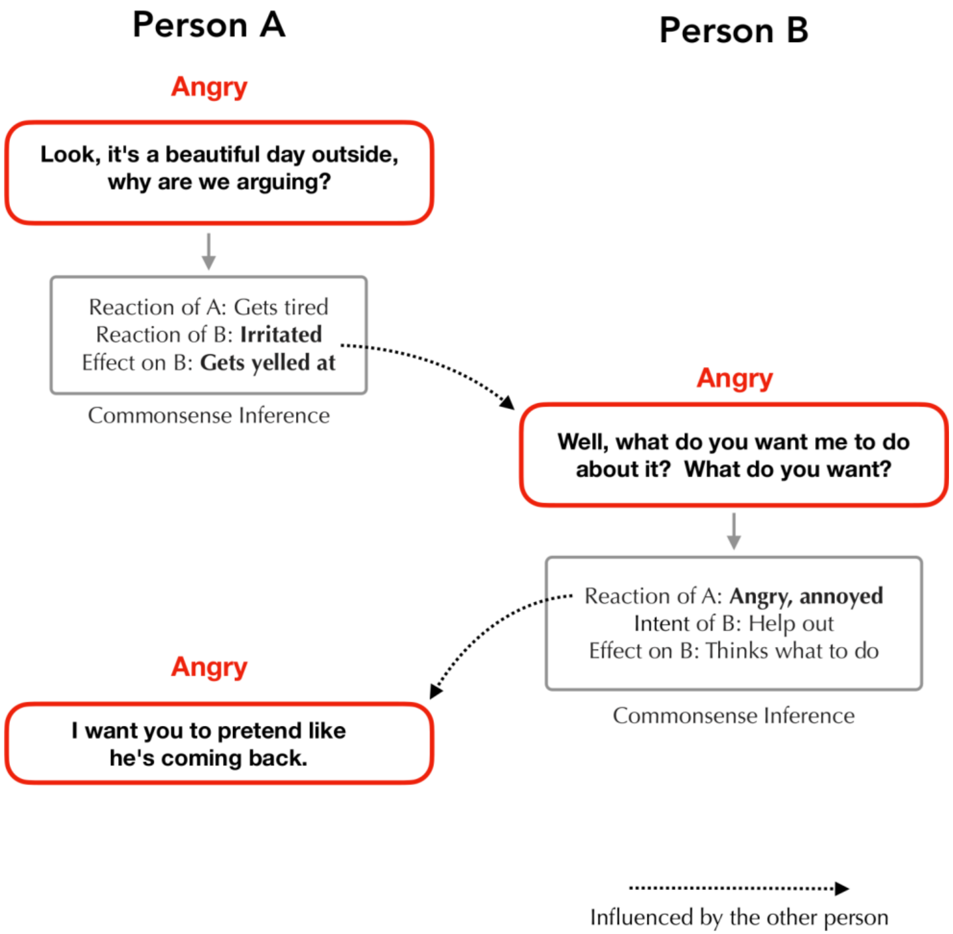
Example: "Person X gives Person Y a compliment"

→ Intend of X: "X wanted to be nice"

→ Reaction of Y: "Y will feel flattered"

"COSMIC: COmmonSense knowledge for eMotion Identification in Conversations", Findings of EMNLP 2020

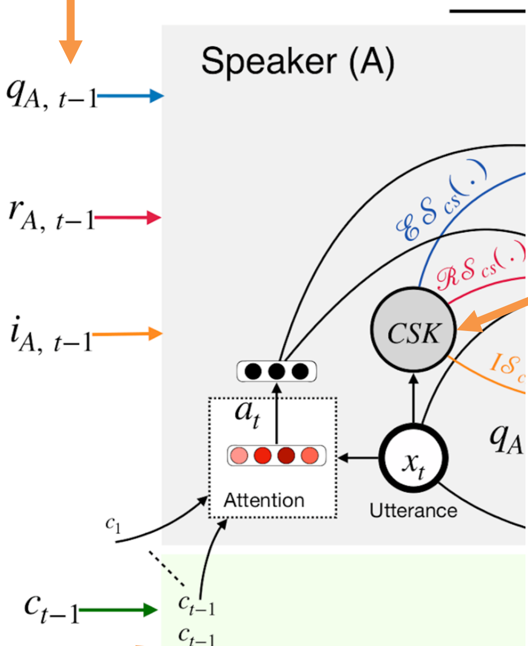
Commonsense and emotion recognition



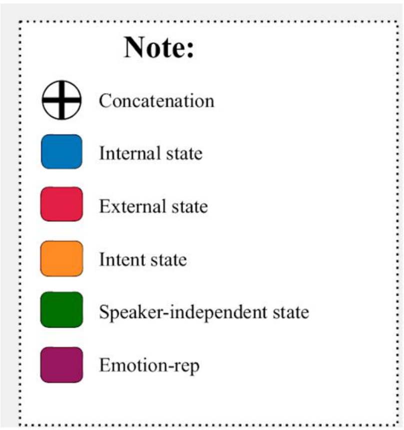
"COSMIC: COmmonSense knowledge for eMotion Identification in Conversations", Findings of EMNLP 2020

Proposed Model (COSMIC)

Previous internal/external/intent state



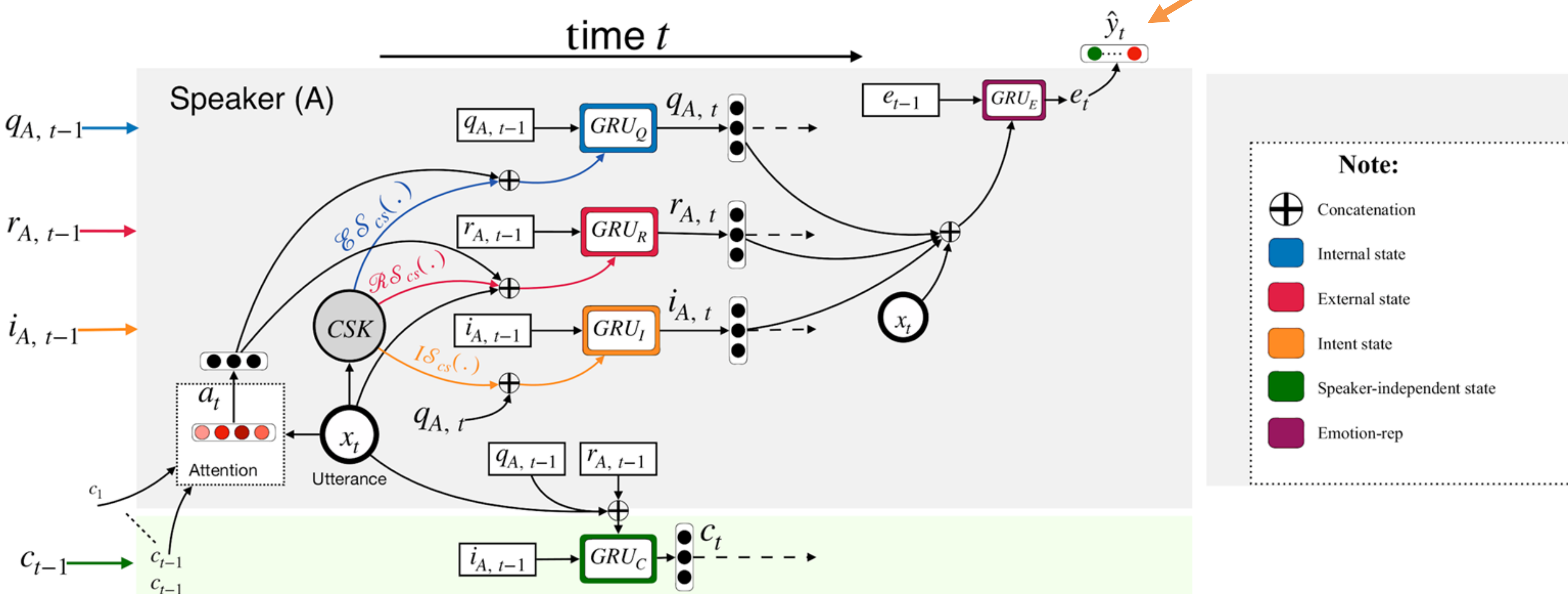
COMET embeddings



Previous dyadic state

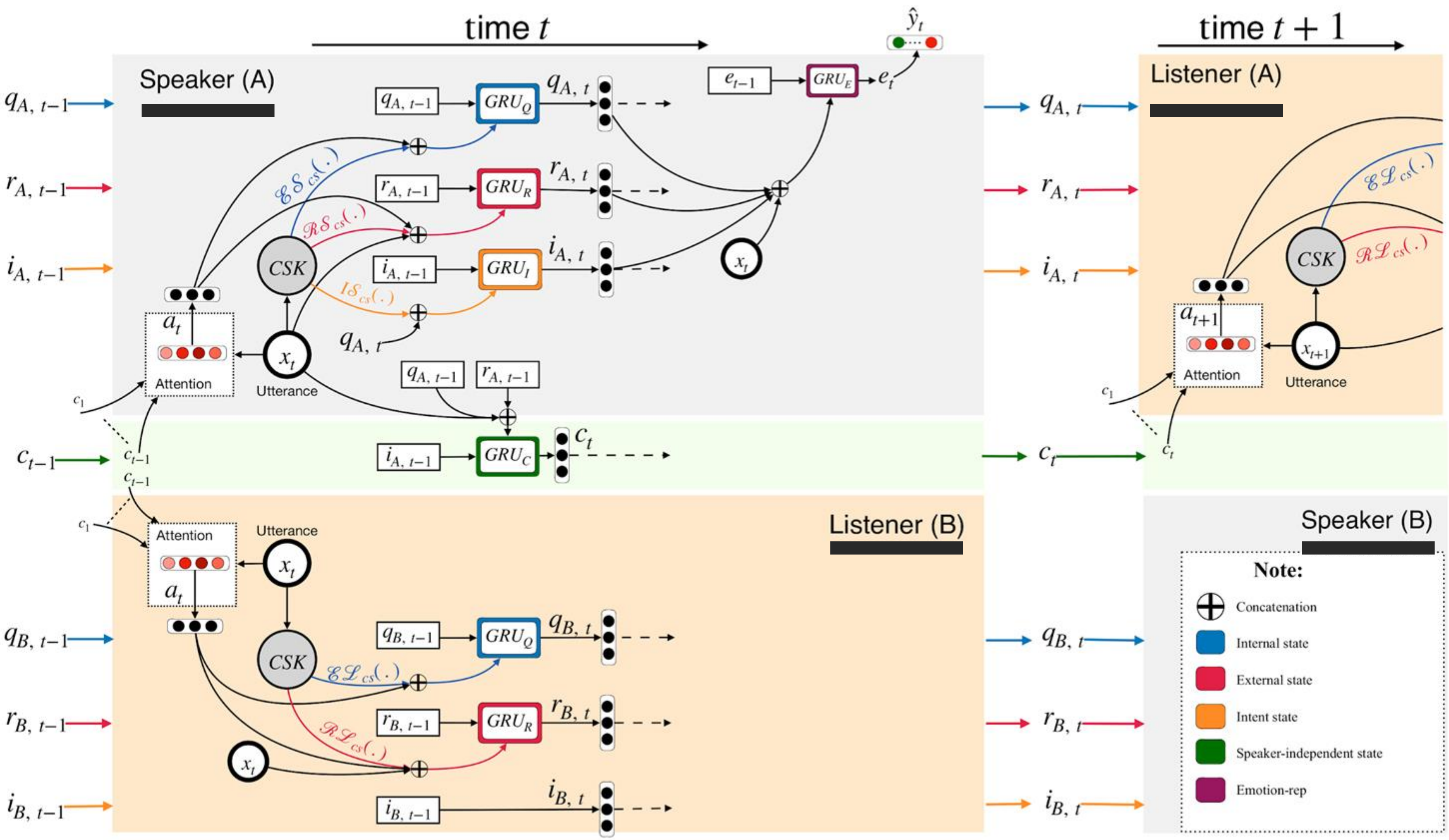
Proposed Model (COSMIC)

Make a prediction



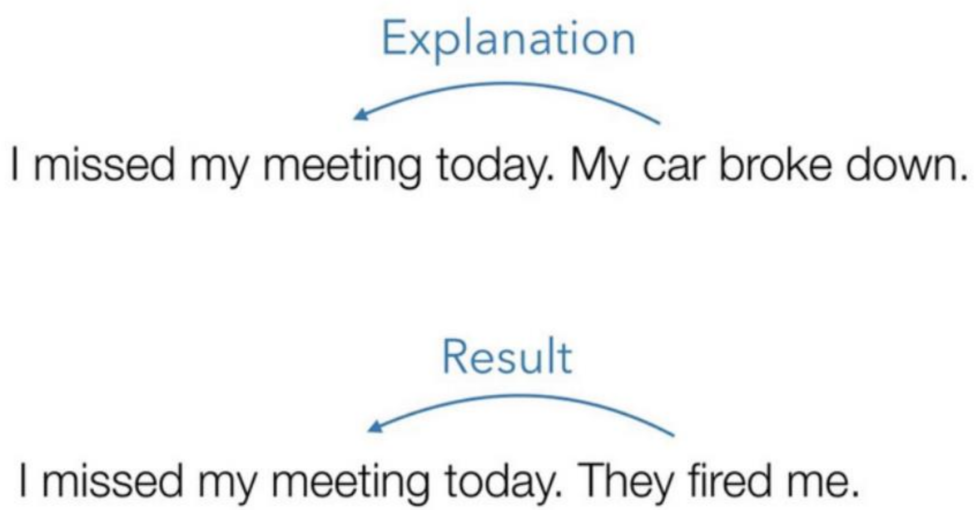
Next dyadic state

Proposed Model (COSMIC)



Coherence and Commonsense

Coherence relations provide information about how the content of discourse units relate to one another.
They have been used to predict **commonsense inference** in text.



Cross-modal Coherence Modeling for Caption Generation

NEW paper

Research task: Coherence relation prediction for imagery and text

Visible: horse and rider jumping a fence.

Meta: horse and rider jumping a fence during a race.

Subjective: the most beautiful horse in the world.

Story: horse competes in the event.

➡ Cross-modal coherence modeling can help systems to recognize that image descriptions can **fulfill different purposes**.

Cross-modal Coherence Modeling for Caption Generation ACL 2020

Cross-modal Coherence Modeling for Caption Generation

New dataset: Coherence relations between image-text pairs are collected, such as captions can be subjective, action oriented, meta, story,…

➡ Image captions are subjective, and several relations can hold concurrently

visible, action, subjective

Photo credit: Shutterstock user yauhenka

Young happy boy swimming in the lake.

10,000 image–text pairs annotated by expert annotators with a high agreement.

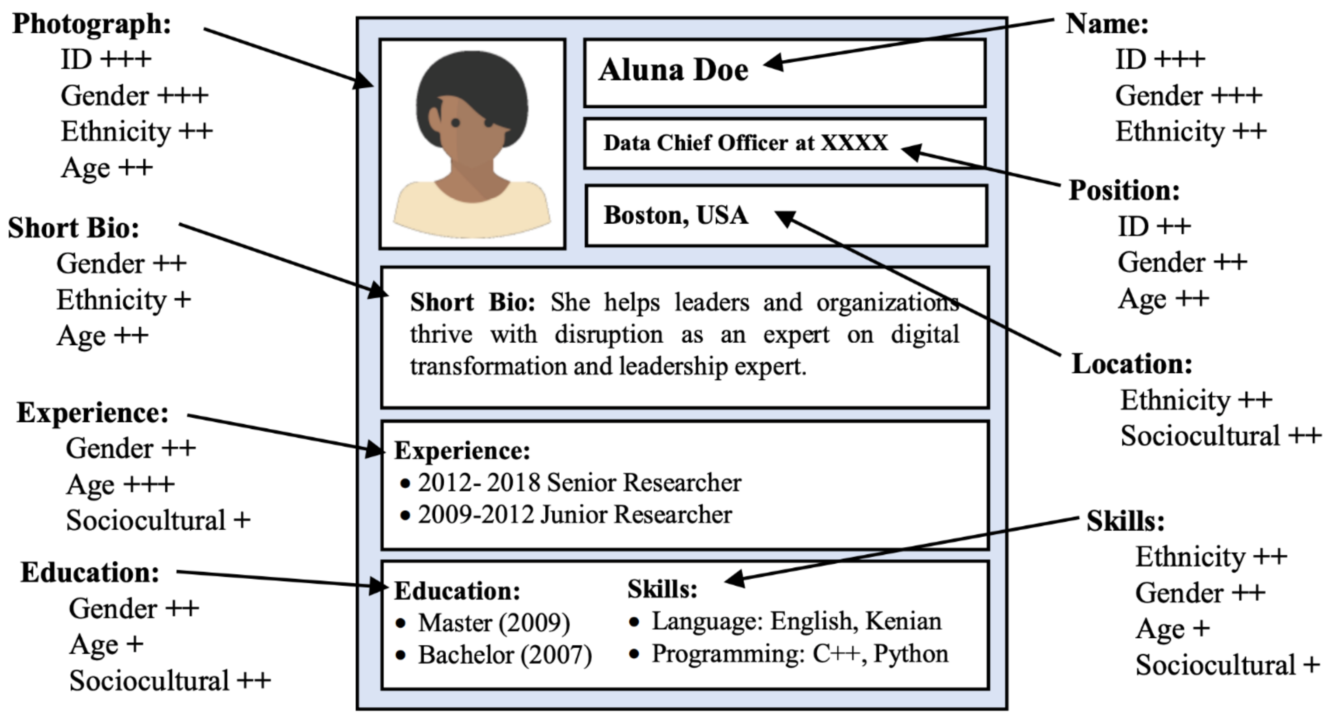
- 5,000 from Conceptual Captions (Sharma et al., 2018)
- 5,000 from machine-authored captions from the state of the art models in 2019

Social Impact – Fairness and Misinformation

Fair Representation Learning

NEW paper

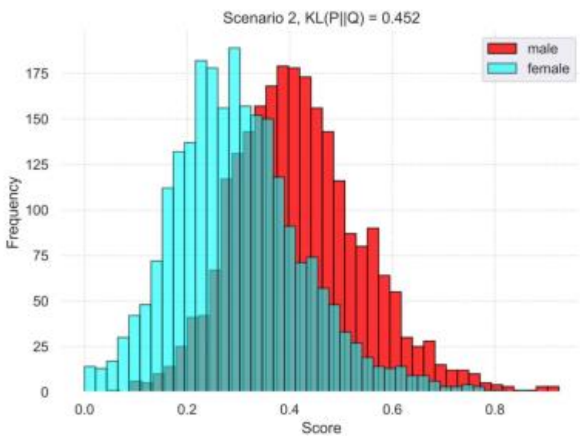
24,000 synthetic resumes to test biases in multimodal prediction

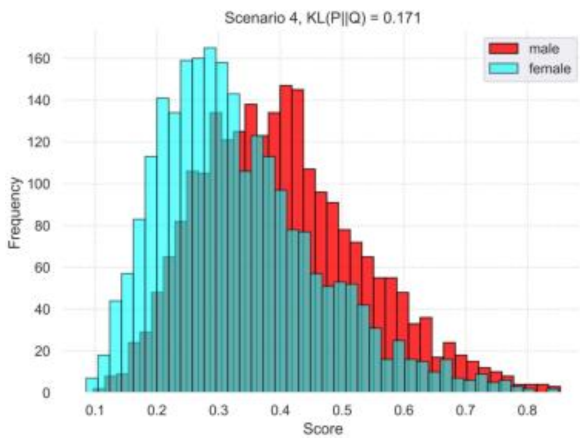


Pena et al., Bias in Multimodal AI: A Testbed for Fair Automatic Recruitment. ICMI 2020

Fair Representation Learning

Finding: Multimodal models reproduce biases present in the training data even if the gender attribute is not explicitly available.





Significant differences in predicted distributions wrt gender and race

Pena et al., Bias in Multimodal AI: A Testbed for Fair Automatic Recruitment. ICMI 2020

Fair Representation Learning

Towards mitigating biases: minimizing both prediction loss and *sensitivity* (the amount of sensitive information in the learned model represented)

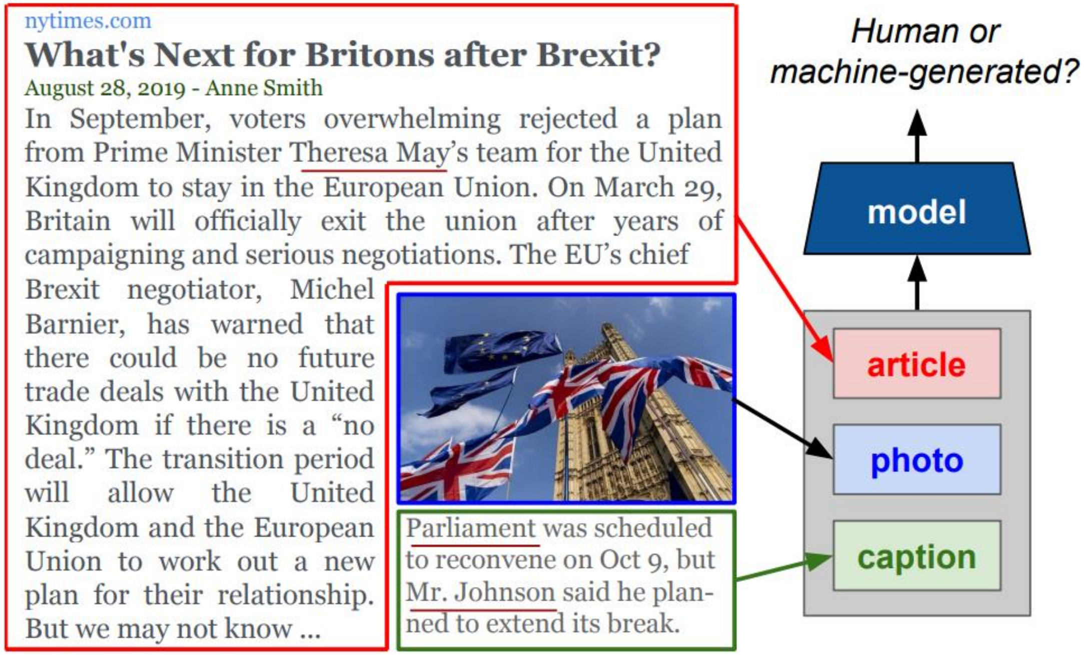
Scenario	Bias	Input Features			Gender		Δ	Ethnicity			Δ
		Merits	Dem	Face	Male	Female		Group 1	Group 2	Group 3	
1	no	yes	yes	no	51%	49%	2%	33%	34%	33%	1%
2	yes	yes	yes	no	87%	13%	74%	90%	9%	1%	89%
3	yes	yes	no	no	50%	50%	0%	32%	34%	34%	2%
4	yes	yes	no	yes	77%	23%	54%	53%	31%	16%	37%
Agnostic	yes	yes	no	yes	50%	50%	0%	35%	30%	35%	5%

Pena et al., Bias in Multimodal AI: A Testbed for Fair Automatic Recruitment. ICMI 2020

Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News

NEW paper

New task: Defending against full news article containing image-caption pairs.



Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News, EMNLP 2020

Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News

New dataset: NeuralNews dataset that contains both human and machine-generated articles with images and captions.

# Sentences in Article	% of Articles		# Imgs	% of Articles
	Real	Generated		
$N \leq 10$	33.7	15.6	1	60.8
$10 < N \leq 40$	54.4	81.5	2	21.0
$N > 40$	11.9	2.9	3	18.2

Proposed model: Propose DIDAN, an effective named entity-based model that serves as a good baseline for defending against neural fake news.

Emotional and Engaging Interactions

Dialogue Act Classification (DAC)

NEW paper

Dialogue act labels:

Greeting, Question, Answer, Statement-Opinion, Statement-Non-Opinion, Apology, Command, Agreement, Disagreement, Acknowledge, Backchannel, and Others

Research questions:

→ Are video+audio helpful for DAC?

→ Are emotions helpful for DAC

→ Is DA helpful for emotion recognition?

"Towards Emotion-aided Multi-modal Dialogue Act Classification", ACL 2020

Emotional Dialogue Act Classification

New dataset: EMO-TyDA which adds 12 most common DAC annotations to two pre-existing datasets (IEMOCAP and MELD)

	IEMOCAP		MELD	
	# Utterance	**# Dialogue**	**# Utterance**	**# Dialogue**
Train	7497	242	7489	831
Test	1879	60	2500	208

"Towards Emotion-aided Multi-modal Dialogue Act Classification"

Emotional Dialogue Act Classification

Example from MELD:

Utterance

1) **Phoebe:** Fine! Then you tell Roger because he was really looking forward to this!

- **Text** : suggests agreement or opinion
- **Audio :** commanding tone
- **Video :** furious

2) **M_1:** That's very amusing indeed.

- **Text** : agreement
- **Audio :** sarcastic tone
- **Video :** slight anger

"Towards Emotion-aided Multi-modal Dialogue Act Classification", ACL 2020

Emotional Dialogue Act Classification

Example from IEMOCAP:

	Utterance	Emotion
1)	**Monica:** I can't leave it! You gouged a hole in my dingy floor.	anger
	DA: disagreement	
2)	**M_2:** Well, you know I appreciate you coming over and talking to me, I mean it definitely helps.	sad
	DA: acknowledge	

"Towards Emotion-aided Multi-modal Dialogue Act Classification"

Image-Chat: Engaging Grounded Conversations

New dataset - Image-Chat: image grounded dialogs where the annotators are given a specific speaking style to follow.



A: Peaceful B: Absentminded	*A: Fearful B: Miserable*	*A: Erratic B: Skeptical*
A: I'm so thankful for this delicious food.	A: I just heard something out there and I have no idea what it was.	A: What is the difference between the forest and the trees? Oh look, dry pavement.
B: What is it called again?	B: It was probably a Wolf coming to eat us because you talk too much.	B: I doubt that's even a forest, it looks like a line of trees.
A: Not sure but fried goodness.	A: I would never go camping in the woods for this very reason.	A: There's probably more lame pavement on the other side!

Shuster, Kurt, et al. "Image-chat: Engaging grounded conversations." ACL 2020

Multi-Lingual Multimodal Grounding

Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding

NEW paper

Task: Follow navigation instructions through a home

Our starting point is in a living room, we're facing towards a long beige sofa, and in front of the sofa there are three glass coffee tables, turn around and exit through the doorway that's in front of you, walk pass the bed that's on your right and then turn left, we're now facing towards another living room, and on the left there's an open door, walk towards that open door enter the bathroom that's in front of you, turn towards the right into the shower area, and that's your destination.

Lessons from prior work
(e.g. Room-to-Room)

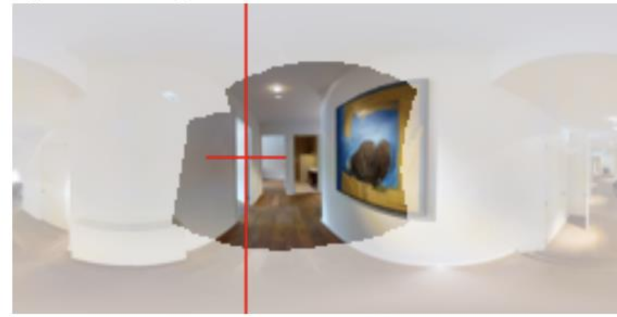
1. R2R's paths were too short to guarantee instruction following vs search
2. R2R's paths had biases that could be learned without vision/language
3. R2R was only in English

Room-Across-Toom: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding
Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge – EMNLP 2020

Room-Across-Room Dataset

Dataset design motivations:

1. High variance in path lengths (avoid length prior informing agents)
2. Paths may be circuitous (test if following directions or finding goal)
3. Uniform coverage of environment viewpoints (avoid instructions collapsing to single referent per room)



...crossing a wall painting which is to your right side, you can see open door enter...

...enter into it. This is a gym room, move forward, walk...

Room-Across-Toom: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding
Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge – EMNLP 2020

Multilingual Statistics

	R2R		RxR								
	en		hi		te		en-IN		en-US		
Phenomenon	p	μ	p	μ	p	μ	p	μ	p	μ	**RxR Example (en-US)**
Reference	100	3.7	100	5.8	100	6.6	100	6.4	100	8.3	...there is **a white chair** and **a table stand**...
Coreference	32	0.5	40	0.4	76	2.9	76	6.4	64	5.3	...hallway with black curtains, towards **that**...
Comparison	4	0.0	0	0.0	4	0.1	4	0.0	8	0.0	...the large archway with the **smaller** archway in...
Sequencing	16	0.2	24	0.2	44	0.6	44	0.5	52	0.9	...the **next** room... turn to see the **next** door...
Allocentric Relation	20	0.2	68	2.1	76	3.2	92	3.4	76	2.4	...a window with a black folding table **under** that...
Egocentric Relation	80	1.2	96	2.9	80	2.3	64	2.8	60	2.3	...chairs on **your right**, closet doors on **your left**.
Imperative	100	4.0	100	5.6	100	6.5	100	8.4	100	6.3	**Do not** go down the stairs. Instead, **look** further...
Direction	100	2.8	96	5.8	96	4.9	100	7.0	96	6.3	...**veer to the left** of the fireplace and you will...
Temporal Condition	28	0.4	32	0.4	36	0.7	44	1.0	52	0.8	Move around the island **until** you come to the...
State Verification	8	0.1	72	1.7	68	1.6	80	2.3	84	3.1	...**you are in** the balcony area facing towards...

P is the % of sentences with a given phenomena vs average # of times within a sentence

Room-Across-Toom: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding
Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge – EMNLP 2020

Multilingual Multimodal Agents

Paths are collected by **G**uides (giving) and **F**ollowers (taking) said paths

1. Is it helpful to train an agent based on both? **Yes**

		Setting			Training	NE ↓			SR ↑			SDTW ↑			NDTW ↑		
Exp.	Method	G	F	X	Pairs (K)	en	hi	te	en	hi	te	en	hi	te	en	hi	te
(1)	Mono	✓			42	10.1	9.7	9.4	25.6	24.8	28.0	20.3	19.7	22.7	41.3	38.8	43.7
(2)	Mono		✓		42	10.3	**9.2**	9.5	23.9	28.0	27.0	18.5	22.7	22.0	37.0	**45.9**	43.9
(3)	Mono	✓	✓		84	**9.8**	**9.2**	**9.1**	**26.1**	**29.6**	**29.8**	**21.0**	**24.0**	**24.2**	**42.4**	45.5	**45.6**

2. Is it helpful to train in multiple languages at the same time? **Ergh, um, no?**

Exp.	Method	G	F	X	Pairs (K)	NE en	NE hi	NE te	SR en	SR hi	SR te	SDTW en	SDTW hi	SDTW te	NDTW en	NDTW hi	NDTW te
(4)	Multi	✓	✓		252	**11.0**	10.9	11.0	**22.2**	**23.0**	23.1	**17.8**	**18.3**	**18.4**	**38.6**	39.2	38.8
(5)	Multi	✓	✓	✓	504	11.5	11.4	11.4	20.0	18.7	20.3	15.9	14.9	16.1	36.3	36.0	36.7
(6)	Multi*	✓	✓		252	**11.0**	**10.7**	**10.7**	21.9	22.6	**23.2**	17.5	18.1	**18.4**	**38.6**	**39.9**	**39.7**
(H)	Human				-	1.32	0.59	0.79	90.4	96.8	94.7	74.3	80.6	76.5	77.7	82.2	79.2

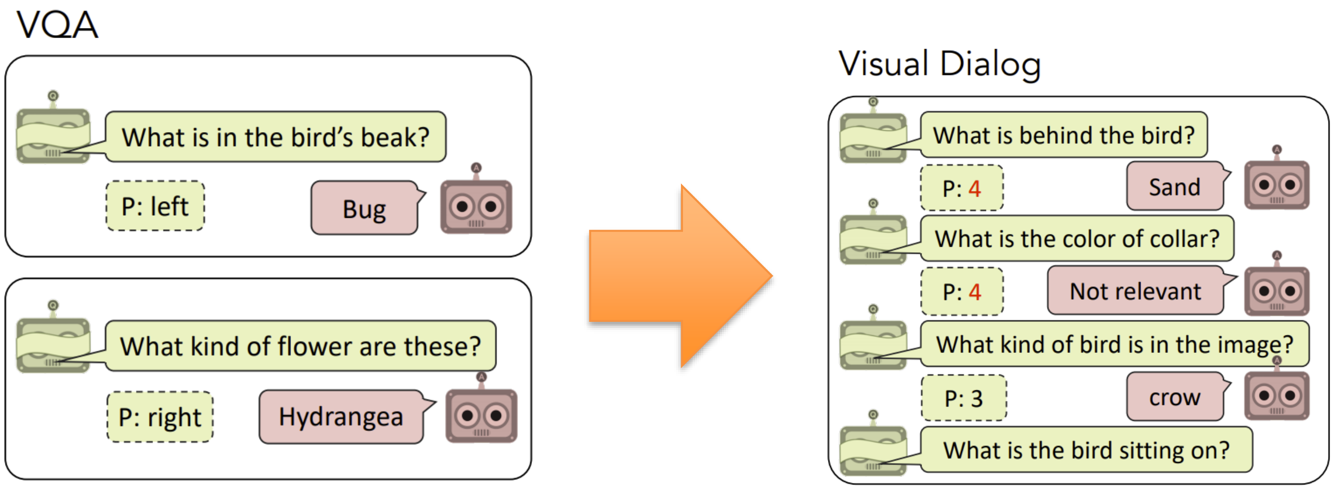
Settings – G: instruction paired with Guide paths, F: instructions paired with Follower paths, X: cross-translated instructions.

Room-Across-Toom: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding
Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge – EMNLP 2020

Dialog without Dialog Data: Learning Visual Dialog Agents from VQA Data

NEW paper

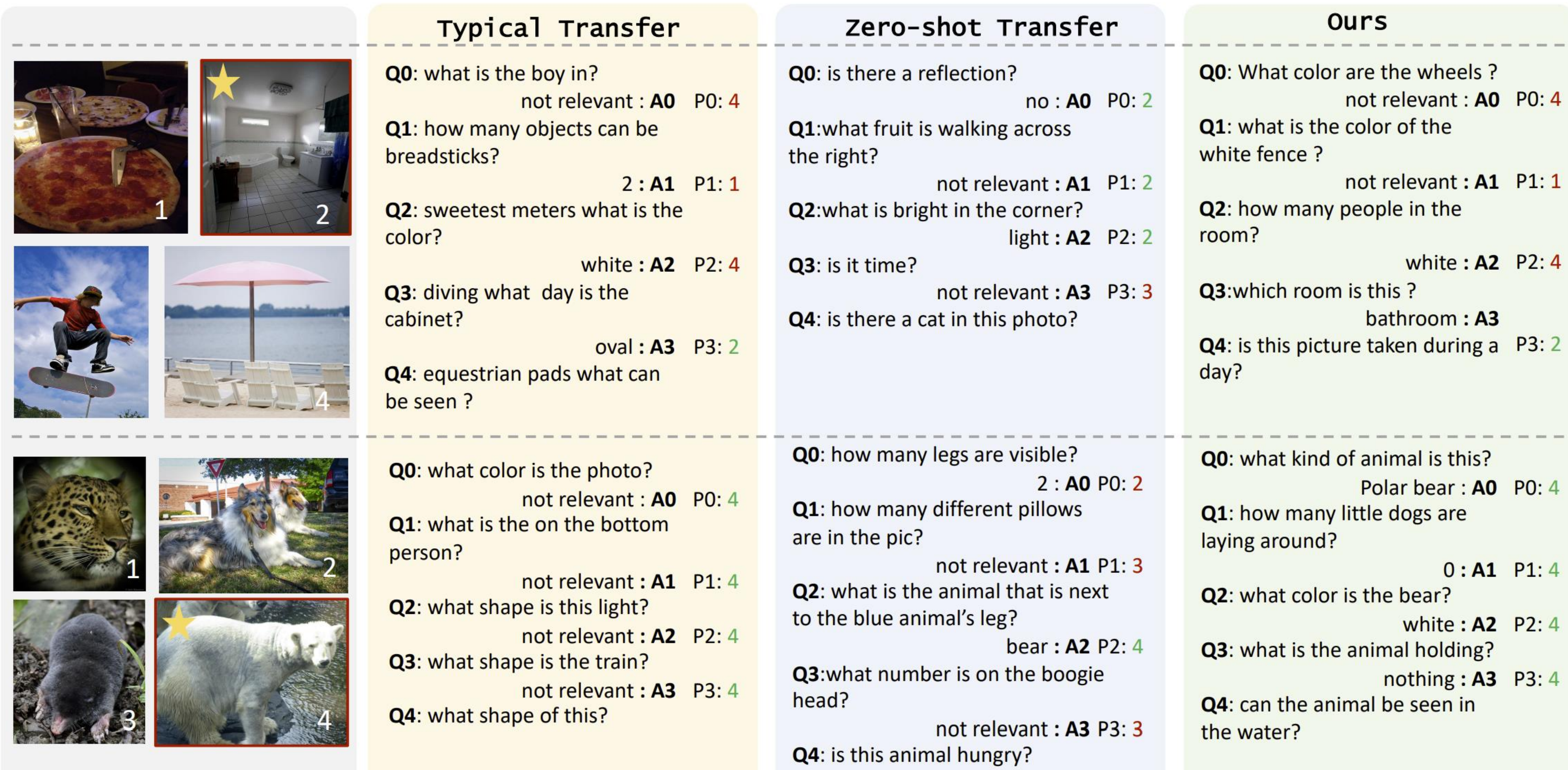
Problem: Can we develop visually-grounded dialog agents from data which does not contain multiple dialogue turns?



Main idea: Try to decouple the question intent from the specific words

Cogswell, Michael, et al. "Dialog without Dialog Data: Learning Visual Dialog Agents from VQA Data."

Dialog without Dialog Data: Learning Visual Dialog Agents from VQA Data



Cogswell, Michael, et al. "Dialog without Dialog Data: Learning Visual Dialog Agents from VQA Data."

References

Abstraction and Logic

- Learning by Abstraction: The Neural State Machine, Neurips 2019
- VQA-LOL: Visual Question Answering under the Lens of Logic, ECCV 2020

Multimodal Reasoning

- Cross-Modality Relevance for Reasoning on Language and Vision ACL 2020
- Multi-step Reasoning via Recurrent Dual Attention for Visual DialogACL 2019

Towards Causal Inference

- Two Causal Principles for Improving Visual Dialog CVPR 2020
- Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing, CVPR 2020

Understanding Multimodal Models

- SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions, CVPR 2020
- What Makes Training Multi-modal Classification Networks Hard?, CVPR 2020

Coherence and Commonsense

- COSMIC: COmmonSense knowledge for eMotion Identification in Conversations , Findings of EMNLP 2020
- Cross-modal Coherence Modeling for Caption Generation ACL 2020

Social Impact – Fairness and Misinformation

- Bias in Multimodal AI: Testbed for Fair Automatic Recruitment, CVPR-W 2020, ICMI 2020
- Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News, EMNLP 2020

Emotional and Engaging Interactions

- Image-Chat: Engaging Grounded Conversations, ACL 2020
- Towards Emotion-aided Multi-modal Dialogue Act Classification, ACL 2020

Multi-Lingual-Multimodal Grounding

- Room-across-Room: Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding -- EMNLP 2020
- Dialog without Dialog Data: Learning Visual Dialog Agents from VQA Data, NeurIPS 2020