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
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.
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?

Learning by Abstraction: The Neural State Machine

Pre-trained an alphabet of concepts (Visual Genome)

Manually grouped by “properties”

Probabilities computed at runtime for each object instance

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?

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

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

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

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.

# Learning by Abstraction: The Neural State Machine

**Content Generalization**

<table>
<thead>
<tr>
<th></th>
<th>training</th>
<th>testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only questions that <strong>do not</strong> refer to any type of <strong>food</strong> or <strong>animal</strong> (do not include any word from these categories)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Structure Generalization**

<table>
<thead>
<tr>
<th></th>
<th>training</th>
<th>testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the <code>&lt;obj&gt;</code> covered by?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is there a <code>&lt;obj&gt;</code> in the image?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the <code>&lt;obj&gt;</code> made of?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What’s the name of the <code>&lt;obj&gt;</code> that is <code>&lt;attr&gt;</code>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Content</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Prior</td>
<td>8.51</td>
<td>14.64</td>
</tr>
<tr>
<td>Local Prior</td>
<td>12.14</td>
<td>18.21</td>
</tr>
<tr>
<td>Vision</td>
<td>17.51</td>
<td>18.68</td>
</tr>
<tr>
<td>Language</td>
<td>21.14</td>
<td>32.88</td>
</tr>
<tr>
<td>Lang+Vis</td>
<td>24.95</td>
<td>36.51</td>
</tr>
<tr>
<td>BottomUp [5]</td>
<td>29.72</td>
<td>41.83</td>
</tr>
<tr>
<td>MAC [40]</td>
<td>31.12</td>
<td>47.27</td>
</tr>
<tr>
<td>NSM</td>
<td><strong>40.24</strong></td>
<td><strong>55.72</strong></td>
</tr>
</tbody>
</table>

---

VQA under the Lens of Logic

VQA under the Lens of Logic

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

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

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

Cross-Modality Relevance for Reasoning on Language and Vision

Visual Question Answering

Text: Where is the child sitting?

- fridge
- arms

Natural Language for Visual Reasoning

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

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

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
Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog

**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**
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.

Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog

Coupling of the dual attentions

Towards Causal Inference
Visual Dialogue Expressed with Causal Graph

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

X → Y
cause → effect

Important assumption: the output of a neural network is the effect of the input (the cause)
Two Causal Principles for Improving Visual Dialog

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.

Two Causal Principles for Improving Visual Dialog

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

**Dataset bias example:**

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.

Two Causal Principles for Improving Visual Dialog

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

**Dataset bias example:**

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

Studying Biases in VQA Models

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…

<table>
<thead>
<tr>
<th>Q: What are the shelves made of?</th>
<th>A: glass</th>
<th>vases removed; A: glass</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+LSTM</td>
<td>glass</td>
<td>wood</td>
</tr>
<tr>
<td>SAAA</td>
<td>glass</td>
<td>metal</td>
</tr>
<tr>
<td>SNMN</td>
<td>glass</td>
<td>metal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q: Are there zebras in the picture?</th>
<th>A: yes</th>
<th>giraffes removed; A: yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+LSTM</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>SAAA</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>SNMN</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q: What sport is he playing?</th>
<th>A: soccer</th>
<th>sports-ball; A: soccer</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+LSTM</td>
<td>soccer</td>
<td>tennis</td>
</tr>
<tr>
<td>SAAA</td>
<td>soccer</td>
<td>tennis</td>
</tr>
<tr>
<td>SNMN</td>
<td>soccer</td>
<td>tennis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q: How many dogs are there?</th>
<th>A: 1</th>
<th>dog removed; A: 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+LSTM</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>SAAA</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SNMN</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."
**Proposed solution:** training the model on original VQA datasets plus synthetic datasets, consisting of images with removed objects.

<table>
<thead>
<tr>
<th>Q: Is there a bowl on the table?</th>
<th>A: no</th>
<th>cup removed; A: no</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CL</strong></td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td><strong>SAAA</strong></td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td><strong>SNMN</strong></td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q: How many people are in the water?</th>
<th>A: 1</th>
<th>person removed; A: 0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CL</strong></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>SAAA</strong></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>SNMN</strong></td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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

Does VQA model have the right "reasoning" of getting the right answer?

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.

SQuINtIng Model

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

What Makes Training Multi-modal Classification Networks Hard?

Adding more modalities should always help?

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Multi-modal</th>
<th>V@1</th>
<th>Best Uni</th>
<th>V@1</th>
<th>Drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinetics</td>
<td>A + RGB</td>
<td>71.4</td>
<td>RGB</td>
<td>72.6</td>
<td>-1.2</td>
</tr>
<tr>
<td></td>
<td>RGB + OF</td>
<td>71.3</td>
<td>RGB</td>
<td>72.6</td>
<td>-1.3</td>
</tr>
<tr>
<td></td>
<td>A + OF</td>
<td>58.3</td>
<td>OF</td>
<td>62.1</td>
<td>-3.8</td>
</tr>
<tr>
<td></td>
<td>A + RGB + OF</td>
<td>70.0</td>
<td>RGB</td>
<td>72.6</td>
<td>-2.6</td>
</tr>
</tbody>
</table>

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)

$\phi_{m_1}$ $\phi_{m_2}$

$\mathcal{L}_{\text{multi}}$

Proposed approach

$\phi_{m_1}$ $\phi_{m_2}$

$w_1 \mathcal{L}_1$ $w_{\text{multi}} \mathcal{L}_{\text{multi}}$ $w_2 \mathcal{L}_2$

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
Emotions are Often Context Dependent

Dialogue:
1) You liked it? You really liked it?
2) Oh, yeah!
3) Which part exactly?
4) The whole thing! Can we go?
5) What about the scene with the kangaroo?
6) I was surprised to see a kangaroo in a world war epic.
7) You fell asleep!
8) Don't go, I’m sorry.

Emotion (Sentiment):
- Surprise (Positive)
- Neutral (Neutral)
- Neutral (Neutral)
- Anger (Negative)

“COSMIC: COnmonSense knowledge for eMotion Identification in Conversations”, Findings of EMNLP 2020
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”
Commonsense and emotion recognition

Person A

Angry

Look, it's a beautiful day outside, why are we arguing?

Reaction of A: Gets tired
Reaction of B: Irritated
Effect on B: Gets yelled at

Commonsense Inference

Person B

Angry

Well, what do you want me to do about it? What do you want?

Reaction of A: Angry, annoyed
Intent of B: Help out
Effect on B: Thinks what to do

Commonsense Inference

I want you to pretend like he's coming back.

Influenced by the other person

“COSMIC: CCommonSense 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.

Explanation

I missed my meeting today. My car broke down.

Result

I missed my meeting today. They fired me.
Cross-modal Coherence Modeling for Caption Generation

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

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

![Image of a boy swimming](image.png)

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

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

*Photo credit: Shutterstock user yashenka*

*Young happy boy swimming in the lake.*
Social Impact – Fairness and Misinformation
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.

![Graphs showing predicted distributions](images)

- 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)

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Bias</th>
<th>Input Features</th>
<th>Gender</th>
<th>Δ</th>
<th>Ethnicity</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Merits Dem Face Male Female</td>
<td></td>
<td></td>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td>1</td>
<td>no</td>
<td>yes yes no</td>
<td>51% 49%</td>
<td>2%</td>
<td>33% 34%</td>
<td>33%</td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
<td>yes yes no</td>
<td>87% 13%</td>
<td>74%</td>
<td>90% 9%</td>
<td>1%</td>
</tr>
<tr>
<td>3</td>
<td>yes</td>
<td>yes no no</td>
<td>50% 50%</td>
<td>0%</td>
<td>32% 34%</td>
<td>34%</td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>yes no yes</td>
<td>77% 23%</td>
<td>54%</td>
<td>53% 31%</td>
<td>16%</td>
</tr>
<tr>
<td>Agnostic</td>
<td>yes</td>
<td>yes no yes</td>
<td>50% 50%</td>
<td>0%</td>
<td>35% 30%</td>
<td>35%</td>
</tr>
</tbody>
</table>

Pena et al., Bias in Multimodal AI: A Testbed for Fair Automatic Recruitment. ICMI 2020
New task: Defending against full news article containing image-caption pairs.

nytimes.com
What's Next for Britons after Brexit?
August 28, 2019 - Anne Smith
In September, voters overwhelming rejected a plan from Prime Minister Theresa May’s team for the United Kingdom to stay in the European Union. On March 29, Britain will officially exit the union after years of campaigning and serious negotiations. The EU’s chief Brexit negotiator, Michel Barnier, has warned that there could be no future trade deals with the United Kingdom if there is a “no deal.” The transition period will allow the United Kingdom and the European Union to work out a new plan for their relationship. But we may not know ...
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.

<table>
<thead>
<tr>
<th># Sentences in Article</th>
<th>% of Articles</th>
<th># Imgs</th>
<th>% of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N \leq 10$</td>
<td>33.7</td>
<td>1</td>
<td>60.8</td>
</tr>
<tr>
<td>$10 &lt; N \leq 40$</td>
<td>54.4</td>
<td>2</td>
<td>21.0</td>
</tr>
<tr>
<td>$N &gt; 40$</td>
<td>11.9</td>
<td>3</td>
<td>18.2</td>
</tr>
</tbody>
</table>

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)

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)

<table>
<thead>
<tr>
<th></th>
<th>IEMOCAP</th>
<th></th>
<th>MELD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Utterance</td>
<td># Dialogue</td>
<td># Utterance</td>
<td># Dialogue</td>
</tr>
<tr>
<td><strong>Train</strong></td>
<td>7497</td>
<td>242</td>
<td>7489</td>
<td>831</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>1879</td>
<td>60</td>
<td>2500</td>
<td>208</td>
</tr>
</tbody>
</table>

“Towards Emotion-aided Multi-modal Dialogue Act Classification”
Emotional Dialogue Act Classification

Example from MELD:

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.

<table>
<thead>
<tr>
<th>A: Peaceful</th>
<th>B: Absentminded</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: I’m so thankful for this delicious food.</td>
<td></td>
</tr>
<tr>
<td>B: What is it called again?</td>
<td>A: Not sure but fried goodness.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A: Fearful</th>
<th>B: Miserable</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: I just heard something out there and I have no idea what it was.</td>
<td>B: It was probably a Wolf coming to eat us because you talk too much.</td>
</tr>
<tr>
<td>A: I would never go camping in the woods for this very reason.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A: Erratic</th>
<th>B: Skeptical</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: What is the difference between the forest and the trees? Oh look, dry pavement.</td>
<td>B: I doubt that’s even a forest, it looks like a line of trees.</td>
</tr>
<tr>
<td>A: There’s probably more lame pavement on the other side!</td>
<td></td>
</tr>
</tbody>
</table>

Multi-Lingual Multimodal Grounding
Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding

Task: Follow navigation instructions through a home

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

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 past 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.

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)
## Multilingual Statistics

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>R2R</th>
<th>R2R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>en</td>
<td>hi</td>
</tr>
<tr>
<td>Reference</td>
<td>100</td>
<td>3.7</td>
</tr>
<tr>
<td>Coreference</td>
<td>32</td>
<td>0.5</td>
</tr>
<tr>
<td>Comparison</td>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td>Sequencing</td>
<td>16</td>
<td>0.2</td>
</tr>
<tr>
<td>Allocentric Relation</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>Egocentric Relation</td>
<td>80</td>
<td>1.2</td>
</tr>
<tr>
<td>Imperative</td>
<td>100</td>
<td>4.0</td>
</tr>
<tr>
<td>Direction</td>
<td>100</td>
<td>2.8</td>
</tr>
<tr>
<td>Temporal Condition</td>
<td>28</td>
<td>0.4</td>
</tr>
<tr>
<td>State Verification</td>
<td>8</td>
<td>0.1</td>
</tr>
</tbody>
</table>

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

Paths are collected by Guides (giving) and Followers (taking) said paths

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

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Method</th>
<th>Setting</th>
<th>Training</th>
<th>NE ↓</th>
<th>SR ↑</th>
<th>SDTW ↑</th>
<th>NDTW ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Mono</td>
<td>✓</td>
<td>42</td>
<td>10.1</td>
<td>25.6</td>
<td>20.3</td>
<td>41.3</td>
</tr>
<tr>
<td>(2)</td>
<td>Mono</td>
<td>✓</td>
<td>42</td>
<td>10.3</td>
<td>23.9</td>
<td>18.5</td>
<td>37.0</td>
</tr>
<tr>
<td>(3)</td>
<td>Mono</td>
<td>✓✓</td>
<td>84</td>
<td>9.8</td>
<td>26.1</td>
<td>21.0</td>
<td>42.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Method</th>
<th>Setting</th>
<th>Training</th>
<th>NE ↓</th>
<th>SR ↑</th>
<th>SDTW ↑</th>
<th>NDTW ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4)</td>
<td>Multi</td>
<td>✓✓</td>
<td>252</td>
<td>11.0</td>
<td>22.2</td>
<td>17.8</td>
<td>38.6</td>
</tr>
<tr>
<td>(5)</td>
<td>Multi</td>
<td>✓✓✓</td>
<td>504</td>
<td>11.5</td>
<td>20.0</td>
<td>15.9</td>
<td>36.3</td>
</tr>
<tr>
<td>(6)</td>
<td>Multi*</td>
<td>✓✓</td>
<td>252</td>
<td>10.7</td>
<td>21.9</td>
<td>17.5</td>
<td>38.6</td>
</tr>
<tr>
<td>(H)</td>
<td>Human</td>
<td>-</td>
<td>1.32</td>
<td>90.4</td>
<td>74.3</td>
<td>77.7</td>
<td>82.2</td>
</tr>
</tbody>
</table>

**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.

## 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."

<table>
<thead>
<tr>
<th><strong>Typical Transfer</strong></th>
<th><strong>Zero-shot Transfer</strong></th>
<th><strong>Ours</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q0:</strong> what is the boy in?</td>
<td><strong>Q0:</strong> is there a reflection?</td>
<td><strong>Q0:</strong> what kind of animal is this?</td>
</tr>
<tr>
<td>not relevant : A0</td>
<td>no : A0</td>
<td>Polar bear : A0</td>
</tr>
<tr>
<td><strong>Q1:</strong> how many objects can be breadsticks?</td>
<td><strong>Q1:</strong> what fruit is walking across the right?</td>
<td><strong>Q1:</strong> how many little dogs are laying around?</td>
</tr>
<tr>
<td>2 : A1</td>
<td>not relevant : A1</td>
<td>0 : A1</td>
</tr>
<tr>
<td><strong>Q2:</strong> sweetest meters what is the color?</td>
<td><strong>Q2:</strong> what is bright in the corner?</td>
<td><strong>Q2:</strong> what color is the bear?</td>
</tr>
<tr>
<td>white : A2</td>
<td>light : A2</td>
<td>white : A2</td>
</tr>
<tr>
<td><strong>Q3:</strong> diving what day is the cabinet?</td>
<td><strong>Q3:</strong> is it time?</td>
<td><strong>Q3:</strong> what is the animal holding?</td>
</tr>
<tr>
<td>oval : A3</td>
<td>not relevant : A3</td>
<td>nothing : A3</td>
</tr>
<tr>
<td><strong>Q4:</strong> equestrian pads what can be seen?</td>
<td><strong>Q4:</strong> is there a cat in this photo?</td>
<td><strong>Q4:</strong> can the animal be seen in the water?</td>
</tr>
</tbody>
</table>

---

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 Dialog ACL 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