Learning by Playing

Devi Parikh
Georgia Tech
People coloring a street in rural Virginia.

“Color College Avenue”, Blacksburg, VA, May 2012
It was a great event! It brought families out, and the whole community together.
Q. What are they coloring the street with?
A. Chalk
AI: What a nice picture! What event was this?

User: “Color College Avenue”. It was a lot of fun!

AI: I am sure it was! Do they do this every year?

User: I wish they would. I don’t think they’ve organized it again since 2012.
“Color College Avenue”, Blacksburg, VA, May 2012
Words

Pictures

“Color College Avenue”, Blacksburg, VA, May 2012
Words

Pictures

Common Sense

“Color College Avenue”, Blacksburg, VA, May 2012
Man in blue wetsuit is surfing on wave
Karpathy and Fei-Fei (Stanford) 2015

A group of young people playing a game of Frisbee
Vinyals et al. (Google) 2015

A car is parked in the middle of nowhere
Kiros et al. (University of Toronto) 2015

A pot of broccoli on a stove.
Fang et al. (Microsoft Research) 2015
A man is rescued from his truck that is hanging dangerously from a bridge.
A man is *rescued* from his truck that is hanging *dangerously* from a bridge.
Learning Common Sense

• Text
  – Reporting bias
Reporting bias in text

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoke</td>
<td>11,577,917</td>
<td>244,458</td>
</tr>
<tr>
<td>laughed</td>
<td>3,904,519</td>
<td>169,347</td>
</tr>
<tr>
<td>murdered</td>
<td>2,843,529</td>
<td>11,284</td>
</tr>
<tr>
<td>inhaled</td>
<td>984,613</td>
<td>4,412</td>
</tr>
<tr>
<td>breathed</td>
<td>725,034</td>
<td>34,912</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>hugged</td>
<td>610,040</td>
<td>10,378</td>
</tr>
<tr>
<td>blinked</td>
<td>390,692</td>
<td>20,624</td>
</tr>
<tr>
<td>was late</td>
<td>368,922</td>
<td>31,168</td>
</tr>
<tr>
<td>exhaled</td>
<td>168,985</td>
<td>3,490</td>
</tr>
<tr>
<td>was punctual</td>
<td>5,045</td>
<td>511</td>
</tr>
</tbody>
</table>

[Gordon et al. 2013]
Reporting bias in text

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoke</td>
<td>11,577,917</td>
<td>244,458</td>
</tr>
<tr>
<td>laughed</td>
<td>3,904,519</td>
<td></td>
</tr>
<tr>
<td>murdered</td>
<td>2,843,529</td>
<td>11,284</td>
</tr>
<tr>
<td>inhaled</td>
<td>984,613</td>
<td>4,412</td>
</tr>
<tr>
<td>breathed</td>
<td>725,034</td>
<td>34,912</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>hugged</td>
<td>610,040</td>
<td>10,378</td>
</tr>
<tr>
<td>was late</td>
<td>390,692</td>
<td>20,624</td>
</tr>
<tr>
<td>exhaled</td>
<td>168,985</td>
<td>3,490</td>
</tr>
<tr>
<td>was punctual</td>
<td>5,045</td>
<td>511</td>
</tr>
</tbody>
</table>

inhale:exhale = 6:1

[Gordon et al. 2013]
Reporting bias in text

murder:exhale = 17:1

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoke</td>
<td>11,577,917</td>
<td>244,458</td>
</tr>
<tr>
<td>laughed</td>
<td>3,904,519</td>
<td>169,347</td>
</tr>
<tr>
<td>murdered</td>
<td>2,843,529</td>
<td>11,284</td>
</tr>
<tr>
<td>inhaled</td>
<td>984,613</td>
<td>4,412</td>
</tr>
<tr>
<td>breathed</td>
<td>725,034</td>
<td>34,912</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>hugged</td>
<td>610,040</td>
<td>10,378</td>
</tr>
<tr>
<td>blinked</td>
<td>390,692</td>
<td>20,624</td>
</tr>
<tr>
<td>was late</td>
<td>368,922</td>
<td>31,168</td>
</tr>
<tr>
<td>exhaled</td>
<td>168,985</td>
<td>3,490</td>
</tr>
<tr>
<td>was punctual</td>
<td>5,045</td>
<td>511</td>
</tr>
</tbody>
</table>

[Gordon et al. 2013]
Reporting bias in text

<table>
<thead>
<tr>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>18,907,427</td>
<td>1,332,154</td>
</tr>
<tr>
<td>Eye(s)</td>
<td>18,455,030</td>
<td>1,090,640</td>
</tr>
<tr>
<td>Arm(s)</td>
<td>6,345,039</td>
<td>458,018</td>
</tr>
<tr>
<td>Ear(s)</td>
<td>3,543,711</td>
<td>230,367</td>
</tr>
<tr>
<td>Brain</td>
<td>3,277,326</td>
<td>260,863</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver</td>
<td>246,937</td>
<td>10,474</td>
</tr>
<tr>
<td>Kidney(s)</td>
<td>183,973</td>
<td>5,014</td>
</tr>
<tr>
<td>Spleen</td>
<td>47,216</td>
<td>1,414</td>
</tr>
<tr>
<td>Pancreas</td>
<td>24,230</td>
<td>1,140</td>
</tr>
<tr>
<td>Gallbladder</td>
<td>17,419</td>
<td>1,556</td>
</tr>
</tbody>
</table>

[Gordon et al. 2013]
Reporting bias in text

People have heads:gallbladders = 1085:1

<table>
<thead>
<tr>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>18,907,427</td>
<td>1,332,154</td>
<td>Liver</td>
<td>246,937</td>
<td>10,474</td>
</tr>
<tr>
<td>Eye(s)</td>
<td>18,455,030</td>
<td>1,090,640</td>
<td>Kidney(s)</td>
<td>183,973</td>
<td>5,014</td>
</tr>
<tr>
<td>Arm(s)</td>
<td>2,488,822</td>
<td>565,123</td>
<td>Pancreas</td>
<td>24,230</td>
<td>1,140</td>
</tr>
<tr>
<td>Ear(s)</td>
<td>3,543,711</td>
<td>230,367</td>
<td>Gallbladder</td>
<td>17,419</td>
<td>1,556</td>
</tr>
<tr>
<td>Brain</td>
<td>3,277,326</td>
<td>260,863</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Gordon et al. 2013]
Learning Common Sense

• Text
  – Reporting bias

• From structure in our visual world?
Two professors converse in front of a blackboard.
Two professors stand in front of a blackboard.
Two professors converse in front of a blackboard.
Challenges

• Lacking visual density
• Annotations are expensive
• Computer vision doesn’t work well enough
Is photorealism necessary?
Create a children's illustration!

Please help create an illustration for a children's story book by creating a realistic scene from the clip art below. Use your imagination! Clipart may be added by double-clicking the clipart in the scene and removed by dragging it off. The clipart may be rotated or flipped and each clipart may only be added once. Please use at least 6 pieces of clipart in each scene. You will be asked to complete 3 different scenes. Press "Next" when finished with the current scene and "Done" when all are finished. Thanks!

Scene 1/3
Mike fights off a bear by giving him a hotdog while Jenny runs away.
1,000 classes of semantically similar scenes:

1,000 classes x 10 scenes per class = 10,000 scenes

[Zitnick and Parikh, CVPR 2013, Oral]
Visual Features
Cloud

Gaze

Smile

Visual Features

Person standing

Basketball

Person sitting

Tree

Cat

Gaze

Slide credit: Larry Zitnick
Visual Features

Which visual features are important for semantic meaning?

Which words correlate with specific visual features?
Generate Scenes

Input: Jenny is catching the ball. Mike is kicking the ball. The table is next to the tree.

Tuples: <<Jenny>,<catch>,<ball>>  <<Mike>,<kick>,<ball>>  <<table>,<be>,<>>

[Zitnick, Parikh and Vanderwende, ICCV 2013]
Learning Fine-grained Interactions

Sentence 1/2: Person 1 is dancing with Person 2

Who is Person 1 in your creation?  ◯ Blonde-haired person  ◯ Brown-haired person

Who is Person 2 in your creation?  ◯ Blonde-haired person  ◯ Brown-haired person

[Antol, Zitnick and Parikh, ECCV 2014]
Learning Fine-grained Interactions

jumping over

holding hands with

dancing with

Train on clipart, test on real
Visual Question Answering (VQA)
Visual Question Answering (VQA)

What is the mustache made of?
Visual Question Answering (VQA)

What is the mustache made of?
Visual Question Answering (VQA)

What is the mustache made of?

AI System

bananas
Visual Question Answering (VQA)

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

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

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

Does it appear to be rainy?
Does this person have 20/20 vision?
Is there a clock ... ?
‘yes’ 98%

Is the man wearing glasses ... ?
‘yes’ 94%

Are the lights on ... ?
‘yes’ 85%

Do you see a ... ?
‘yes’ 87%

Language Bias

Slide credit: Yash Goyal and Peng Zhang
Removing Language Priors

Scene 1/3 - Also need at least: 1 person

**Question** Is there a place to sit other than the floor?

**Answer** yes
Removing Language Priors

Question: Is the girl walking the bike?

Answer: No
Answer: Yes

complementary scenes

[Zhang, Goyal, Summers-Stay, Batra, Parikh, CVPR 2016]
## Classifying a pair of complementary scenes

<table>
<thead>
<tr>
<th>Training set</th>
<th>Unbalanced</th>
<th>Balanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind (no image features)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holistic image features</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Answering Binary Questions

Tuple: <girl, walking, bike>

Question: Is the girl walking the bike?

[Zhang, Goyal, Summers-Stay, Batra, Parikh, CVPR 2016]

Slide credit: Yash Goyal and Peng Zhang
Classifying a pair of complementary scenes

<table>
<thead>
<tr>
<th>Training set</th>
<th>Unbalanced</th>
<th>Balanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind (no image features)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Holistic image features</td>
<td>03.20</td>
<td>23.13</td>
</tr>
<tr>
<td>Attention-based image features</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abstract Scenes

• Learning by playing

• Fully annotated visual data

• Allow full control over the distribution and density of data
  – to learn from
  – to evaluate on
Commonsense Tasks

• Text-based tasks
Key idea

• Imagine the scene behind the text
• Reason about the visual interpretation of the text, not just the text alone
Commonsense Tasks

- Assess plausibility of relations
  - man holds meal
  - tree grows in table

[Vedantam, Lin, Batra, Zitnick, and Parikh, ICCV 2015]
Fill-in-the-blank:

Mike is having lunch when he sees a bear.

A. Mike orders a pizza.
B. Mike hugs the bear.
C. Bears are mammals.
D. Mike tries to hide.
Approach: Imagination

Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
There is a tree near a table.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
Jenny is standing dangerously on the swing.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.

Imagined scenes need not be photorealistic but rich in semantics.
Approach: Imagination

- Clipart Visual World
  
  [CVPR 2013]

  - Two children playing in the park
  - 58 objects
  - 7 poses and 5 expressions
Approach: Imagination

- Scene generation given description [ICCV 2013]

There is a tree near a table. Mike is wearing a blue cap. Mike is telling Jenny to get off the swing.
Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples

\[
\begin{align*}
&<\text{Tree, near table}> \\
&<\text{Mike, wear, cap}> \\
&<\text{Mike, tell, get}> <\text{Jenny, get off, swing}>
\end{align*}
\]
Approach: Imagination

• Scene generation given description [ICCV 2013]
• Semantic parsing into tuples
• Scene generation

Conditional Random Field (CRF)

\[ p(\text{objects} | \text{tuples}) \]
Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

Which objects are present:

- Mike, wear, cap
- Mike, tell, get<br>Jenny, get off, swing

Slide credit: Xiao Lin
Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

Where objects are

Slide credit: Xiao Lin
Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

What are the poses and expressions

Play with

Fight
A.
There is a tree near a table.
B.
The brown dog is standing next to Mike.
C.
The sun is in the sky.
D.
Jenny is standing dangerously on the swing.

__________.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.
Jenny is standing dangerously on the swing. Mike is wearing a blue cap. Mike is telling Jenny to get off the swing.

There is a tree near a table. Mike is wearing a blue cap. Mike is telling Jenny to get off the swing.

\[ w^T \phi_i^{gt} \geq w^T \phi_i^j + 1 \]

Ranking Support Vector Machine (Ranking SVM)

Slide credit: Xiao Lin
## Results

<table>
<thead>
<tr>
<th></th>
<th>Fill-in-the-blanks (FITB) Accuracy (+/- ~0.15)</th>
<th>Visual Paraphrasing (VP) AP (+/- ~0.02)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>25.00</td>
<td>33.33</td>
</tr>
</tbody>
</table>

[Lin and Parikh, CVPR 2015]
## Results

Given *any* tuple, can assess its plausibility

<table>
<thead>
<tr>
<th></th>
<th>Average Precision</th>
<th>Rank Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text alone</td>
<td>72.1</td>
<td>0.488</td>
</tr>
<tr>
<td>Visual alone</td>
<td>68.3</td>
<td>0.461</td>
</tr>
<tr>
<td>Text + visual</td>
<td>73.5</td>
<td>0.504</td>
</tr>
</tbody>
</table>

[Vedantam, Lin, Batra, Zitnick, and Parikh, ICCV 2015]
Visual word2vec

- Learn word embeddings that respect visual (as well as textual) similarity

![Diagram showing visual and textual similarities in word embeddings](image)

[w2v : farther]
- eating
- stares at

[vis-w2v : closer]
- eating
- stares at

Word Embedding

girl
- eating
- ice cream

girl
- stares at
- ice cream

[Kottur, Vedantam and Parikh, CVPR 2016]
Understanding Visual Humor

[Chandrasekaran, Kalyan, Antol, Bansal, Batra, Zitnick, and Parikh, CVPR 2016]
Task 1: Rating humor
Task 2: Remove humor
Task 2: Add humor
Dataset: Abstract Visual Humor (AVH)

Funny

Not funny

Slide credit: Arjun Chandrasekaran
Dataset: Funny Object Replaced (FOR)
Dataset: Funny Object Replaced (FOR)
Dataset: Funny Object Replaced (FOR)
Funny to unfunny
Funny to unfunny
Funny to unfunny
Unfunny to funny
Unfunny to funny
Human evaluation

Humor suppressor

Which scene is **LESS** funny?

5%

95%

Slide credit: Arjun Chandrasekaran
Human evaluation

Humor inducer

Algorithm

Not funny

Slide credit: Arjun Chandrasekaran
Human evaluation

Humor inducer

Not funny

28%

Algorithm

Which scene is MORE funny?

Human

72%

Slide credit: Arjun Chandrasekaran
Funniest scene as per our algorithm

Slide credit: Arjun Chandrasekaran
“This terrified woman's home is being invaded by mice as the cat sleeps.”
“The man is about to trip on his child's car and spill wine on his wife.”
Visual Abstraction For...

- Studying mappings between images and text [CVPR 2013, ICCV 2013]
- Zero-shot learning [ECCV 2014]
- Studying
  - Image memorability [PAMI 2016]
  - Image specificity [CVPR 2015]
  - Visual humor [CVPR 2016]
Visual Abstraction For...

- Studying mappings between images and text [CVPR 2013, ICCV 2013]
- Zero-shot learning [ECCV 2014]
- Studying
  - Image memorability
  - Image specificity
  - Visual humor [CVPR 2016]
- Learning common sense knowledge
  [CVPR 2015, ICCV 2015, CVPR 2016]
- Rich annotation modality
  - Ask for descriptions
  - Ask for scenes
  - Show scene and ask for description
  - Perturb a scene and ask for descriptions
  - ...

Study high-level image understanding tasks without waiting for lower-level vision tasks to be solved

Future work: Learning by “playing”
50k scenes, captions, QAs: available online!
Thank you.