Welcome
2017 Faculty Summit
Attendees

Faculty Summit 2017
microsoftfacultysummit.com

Microsoft Research
Microsoft.com/research

Facebook
@microsoftresearch

Twitter
@MSFTResearch
#FacSumm
#EdgeofAI
On Human Intellect and Machine Failures
Towards Troubleshooting Integrative Machine Learning Systems

Besmira Nushi
Ece Kamar
Eric Horvitz
Donald Kossmann
Exciting times
What happens in the real world?

Google apologises for Photos app’s racist blunder

7 hours ago | Technology

Skyscrapers, Airplanes, Cars, Bikes, Gorillas, Graduation

diir noir avec banan @jacky.aline - Jun 29
Google Photos, y'all. My friend’s not a gorilla.

Mr Alcine tweeted Google about the fact its app had misclassified his photo
What happens in the real world?
Troubleshooting integrative ML systems

Component 1  ---  Component 2  ---  Component 3

ML I/O  ---  ML I/O  ---  system output

I/O

I/O

I/O
Challenge

Possible fixes for each component

Where to invest development effort for biggest impact?

Limited development time
Previous work on troubleshooting
Case study: Image Captioning System

A man flying through the air on a snowboard.

#1

- A man flying through the air on a snowboard.
- A man riding skis on a snowy mountain.

Visual Detector    Language Model    Caption Reranker

I/O                      I/O

snowboard, 0.96
snow, 0.94
man, 0.89
mountain, 0.87
skis, 0.71

...
<table>
<thead>
<tr>
<th>VISUAL DETECTOR</th>
<th>LANGUAGE MODEL</th>
<th>CAPTION RERANKER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. teddy 0.92</td>
<td>1. A teddy bear.</td>
<td>1. A blender sitting on top of a cake.</td>
</tr>
<tr>
<td>2. on 0.92</td>
<td>2. A stuffed bear.</td>
<td>2. A teddy bear in front of a birthday cake.</td>
</tr>
<tr>
<td>3. cake 0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. bear 0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. stuffed 0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. blender 0.57</td>
<td>108. A blender sitting on top of a cake.</td>
<td>3. A cake sitting on top of a blender.</td>
</tr>
</tbody>
</table>
Problem characterization

Continuous quality measures
How to quantify the effects of the component quality continuum on the overall system quality?

Complex component entanglement
How to disentangle the effects of the component failures on the overall system quality?

Non-monotonic error propagation
How to detect whether a component improvement will improve the overall system?
Human-assisted troubleshooting methodology

Component 1 \(\rightarrow\) Component 2 \(\rightarrow\) Component 3

ML I/O \(\rightarrow\) ML I/O

System output
Human-assisted troubleshooting methodology
Human-assisted troubleshooting methodology

Component 1  ML I/O  Component 2  ML I/O  Component 3  system output

[Diagram showing the process of human-assisted troubleshooting with components and system output.]
Human-assisted troubleshooting methodology

Component 1

Component 2

Component 3

ML I/O

ML I/O

system output

Evaluation

Failures

Fixes

Human-aided intervention
Fix workflow execution
Fix workflow execution

Component 1

Component 2

Component 3

ML I/O

system output
Troubleshooting outcomes

Detecting non-monotonic error

Quantifying component impact

Disentangling joint component impact
Case study: Image Captioning System

snowboard, 0.96
snow, 0.94
man, 0.89
mountain, 0.87
skis, 0.71

Visual Detector

Language Model

Caption Reranker

…

A man flying through the air on a snowboard.

A man flying through the air on a snowboard...

A man riding skis on a snowy mountain.

#1

A man flying through the air on a snowboard.
Fixes for the Image Captioning System

Fixes:
- Add object
- Remove object
- Add activity
- Remove activity
- Fix language
- Fix commonsense
- Rerank

System output

Evaluation

Failures

Fixes
Example of Visual Detector Fix

Before fix 😞
A blue and yellow kite.

After fix 😊
A blue umbrella in a field of flowers.
Fix impact summary

- No fix: 57.8% Satisfactory
- Visual Detector
- Language Model
- Caption Reranker
- All fixes
Fix impact summary

<table>
<thead>
<tr>
<th></th>
<th>%Satisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>No fix</td>
<td>57.8</td>
</tr>
<tr>
<td>Visual Detector</td>
<td>68.0</td>
</tr>
<tr>
<td>Language Model</td>
<td>59.3</td>
</tr>
<tr>
<td>Caption Reranker</td>
<td>73.6</td>
</tr>
<tr>
<td>All fixes</td>
<td></td>
</tr>
</tbody>
</table>

Most promising fix
Fix impact summary

Improvement upper bound

Non-monotonic error

Most promising fix
Non-monotonic error propagation

Visual Detector
- teddy 0.92
- computer 0.91
- bear 0.90
- wearing 0.87
- keyboard 0.84
- glasses 0.63

Fixed Visual Detector
- teddy 1.0
- bear 1.0
- wearing 1.0
- keyboard 1.0
- glasses 1.0

1. A teddy bear sitting on top of a computer.

1. A person wearing glasses and holding a teddy bear sitting on top of a keyboard.
Fix impact summary

- Non-monotonic error
- Most promising fix

Improvement upper bound

<table>
<thead>
<tr>
<th></th>
<th>%Satisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>No fix</td>
<td>57.8</td>
</tr>
<tr>
<td>Visual Detector</td>
<td>68.0</td>
</tr>
<tr>
<td>Language Model</td>
<td>59.3</td>
</tr>
<tr>
<td>Caption Reranker</td>
<td>73.6</td>
</tr>
<tr>
<td>All fixes</td>
<td>86.9</td>
</tr>
</tbody>
</table>
Alternative use cases

- Component prototyping
  - Question mark to question

- Architectural changes
  - Q to Q' with ΔQ

- Imperfect fixes
  - Q' to Q'' with ΔQ
  - Noise
Conclusions

Our job does not end with system development.
Continuous monitoring, evaluation, improvements.
Generalizable troubleshooting frameworks are needed.
Human input in troubleshooting

**Crowd worker**
- Reusable tasks
- Clear user interfaces
- Training

**End user**
- Collaboration
- Feedback
- Adjustment
- Interpretation

**System designer**
- Visualization
- Domain analysis
- Interpretation
- Testing
Backup slides
Fix implementation

**Visual Detector**
- Add object
- Remove object
- Add activity
- Remove activity

**Language Model**
- Fix language
- Fix commonsense

**Caption Reranker**
- Rerank

**Visual Detector**
- Improve precision
- Improve recall
- Adjust list size
- Saliency aware

**Language Model**
- Prune bad language
- Prune non csense
- External training

**Caption Reranker**
- Similarity scores
Current system evaluation

Evaluation dataset
1000 images

Satisfactory dataset
578 images (general: 4-5)

Unsatisfactory dataset
422 images (general: 1-3)

Human Satisfaction Scores

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (1-5)</td>
<td>3.674</td>
<td>4.474</td>
<td>2.579</td>
</tr>
<tr>
<td>Detail (1-5)</td>
<td>3.563</td>
<td>4.265</td>
<td>2.601</td>
</tr>
<tr>
<td>Language (1-5)</td>
<td>4.509</td>
<td>4.693</td>
<td>4.256</td>
</tr>
<tr>
<td>Commonsense (0-1)</td>
<td>0.957</td>
<td>1.000</td>
<td>0.898</td>
</tr>
<tr>
<td>General (1-5)</td>
<td>3.517</td>
<td>4.306</td>
<td>2.437</td>
</tr>
<tr>
<td>%Satisfactory (0-1)</td>
<td>57.8%</td>
<td>100%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Satisfactory VS. Unsatisfactory

% Satisfactory

- No fix: 0%
- Visual Detector: 83.3%
- Language Model: 95.8%
- Caption Reranker: 94.5%
- All Fixes: 74.6%
Thank you