Recent progress towards XAI at UC Berkeley

XAI Workshop, ICCV 2019
Prof. Trevor Darrell
The usual XAI story:
Despite conventional wisdom, adding explainability to deep AI models does not decrease their accuracy, and can even improve it.

We don’t need to choose between explainability or high accuracy, can have both!
Inverting the Accuracy-Explainability Curve

DNN XAI systems can lead to better-performing, more explainable models:

1. **Explanations-as-additional-loss**: adding the “show your work” and “right for the right reasons” constraint.

2. **Explanations-allow-advice**: XAI systems transduce DNN states to natural language; reversing this, we can create “Advisable AI” and refine a model via language guidance rather than additional labeled examples.

3. **Explanations-reveal-model-uncertainty**: in a human-in-the-loop retrieval system, explanations let human operators more accurately judge when they should accept suggestions from an XAI teammate.
Today

• Multi-step Saliency via Compositional NMNs
• Fine-grained Textual Explanations
• From Explainable to “Advisable” Driving Models
Attentive XAIs do not decrease recognition accuracy

RISE\(^1\) XAI provides saliency explanations to AI models without affecting their accuracy.

Why does the AI system think these two photos are similar?

SANE\(^2\): When enforcing explainability, attribute recognition performance improves by 2-3\% mAP on two diverse datasets.


Salience for Introspection

Goldfish

- RISE probes black-box CNN models with randomly masked instances of an image to find class-specific evidence
RISE can explain different categories

Explanation for **Sheep**

Explanation for **Cow**

RISE: Randomized Input Sampling for Explanation of Black-box Models, Petsiuk, Das, Saenko, BMVC 2018
RISE: Randomized Input Sampling for Explanation

Neural network prediction: 
**solar farm**: 63%, **shopping mall**: 23%

RISE Explanation for **solar farm**

RISE Explanation for **shopping mall**

Image from the FMoW dataset

Increasing importance
what number of other objects are there of the same size as the gray sphere?

predicted answer: "5"
what number of other objects are there of the same size as the gray sphere?
Multi-step introspection / transparent reasoning

What number of other objects are there of the same size as the gray sphere?

Reasoning

Step 1

look_for("gray sphere")
Multi-step introspection / transparent reasoning

Reasoning
Step 1

look_for("gray sphere")

Reasoning
Step 2

related_by("size")

what number of other objects are there of the same size as the gray sphere?
Multi-step introspection / transparent reasoning

What number of other objects are there of the same size as the gray sphere?

Reasoning

Step 1

look_for("gray sphere")

Step 2

related_by("size")

Step 3

answer("number", "other objects")

predicted answer: "5"  
true answer: "5"
Neural module networks

Example predictions on Visual Question Answering (VQA)
question = “There is a small gray block; are there any spheres to the left of it?”

Example predictions on Referential Expression Grounding (REF)
referential expression = “the cyan thing that is made of the same material as the yellow object”
Monolithic Networks for Visual Question Answering

What is this?

a cat
Monolithic Networks for Visual Question Answering

Monolithic Networks
✓ Work well on simple questions

What is this?

a cat
Monolithic Networks for Visual Question Answering

Monolithic Networks
✓ Work well on simple questions
✗ Challenging for questions requiring *compositional reasoning*
✗ Limited interpretability

*What color is the thing with the same size as the blue cylinder?*
Compositionality in Reasoning

• Generalization to complicated unseen reasoning structure of seen operations (relations)

- How many objects are the either green rubber object or blue cubes?
- Is there a big brown object of the same size as the green thing?
- How many other things are the same size as the yellow rubber ball?
Compositional Inference with Modules

What color is the thing with the same size as the blue cylinder?

```python
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```
Compositional Inference with Modules

What color is the thing with the same size as the blue cylinder?

```
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```
Compositional Inference with Modules

What color is the thing with the same size as the blue cylinder?

```python
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```
Compositional Inference with **Modules**

What color is the thing with the same size as the blue cylinder?

output answer: “gray”

- **Predict** a discrete *execution graph* of modules to answer complex natural language questions
Neural Module Networks (NMNs)

What color is the thing with the same size as the blue cylinder?

```
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```
Dynamic and Reusable Modules

What color is the thing with the same size as the blue cylinder?

- find blue cylinder → compare size → describe color
  - gray

How many things are the same size as the ball?

- find ball → compare size → count
  - four
Dynamic and Reusable Modules

What color is the thing with the same size as the blue cylinder?

gray

How many things are the same size as the ball?

four
End-to-End Module Networks (N2NMN)

How many other things are the same size as the ball?
End-to-End Module Networks (N2NMN)

How many other things are the same size as the ball?

four
End-to-End Module Networks (N2NMN)

How many other things are the same size as the ball?
End-to-End Module Networks (N2NMN)

How many other things are the \textit{same size} as the \textit{ball}?
End-to-End Module Networks (N2NMN)

How many other things are the same size as the ball?
End-to-End Module Networks (N2NMN)

How many other things are the same size as the ball?
End-to-End Module Networks (N2NMN)

How many other things are the same size as the ball?
End-to-End Module Networks (N2NMN)

How many other things are the same size as the ball?

Module Network

find
ball
relocate
same size
count
How many

four
End-to-End Module Networks (N2NMN)

Question RNN encoder

decoder RNN with attention

Layout policy

find

relocate

count

Network builder

find

ball

same size

How many

How many other things are the same size as the ball?

Module Network

four
End-to-End Module Networks (N2NMN)

In this work, we simultaneously learn “what” and “how” end-to-end.
Overview of N2NMN

**Layout policy:** *what tasks to do*

question -> *seq2seq RNN* -> module layout

layout = [Find, Relocate, Count]

**Neural modules:** *how to do them*

layout -> *dynamic networks* -> answer

answer = "4"
Question

“How many other things are of the same size as the green matte ball?”

Translate questions to layout tokens (similar to machine translation)

layout = [find, relocate, count]

assemble dynamic networks

\[ \text{count}(\text{relocate}(\text{find}())) \]

Textual Attention in seq2seq RNNs
Module Networks

• Modules can be added as needed for a given problem

<table>
<thead>
<tr>
<th>Module name</th>
<th>Att-inputs</th>
<th>Features</th>
<th>Output</th>
<th>Implementation details</th>
</tr>
</thead>
<tbody>
<tr>
<td>find</td>
<td>a</td>
<td>$x_{vis}, x_{tst}$</td>
<td>att</td>
<td>$a_{out} = \text{conv}<em>2(\text{conv}<em>1(x</em>{vis}) \odot W</em>{x_{tst}})$</td>
</tr>
<tr>
<td>relocate</td>
<td>a, $a_1, a_2$</td>
<td>$x_{vis}, x_{tst}$</td>
<td>att</td>
<td>$a_{out} = \text{conv}<em>2(\text{conv}<em>1(x</em>{vis}) \odot W_1 \text{sum}(a \odot x</em>{vis}) \odot W_2 x_{tst})$</td>
</tr>
<tr>
<td>and</td>
<td>(none)</td>
<td>$x_{vis}, x_{tst}$</td>
<td>att</td>
<td>$a_{out} = \text{minimum}(a_1, a_2)$</td>
</tr>
<tr>
<td>or</td>
<td>(none)</td>
<td>$x_{vis}, x_{tst}$</td>
<td>att</td>
<td>$a_{out} = \text{maximum}(a_1, a_2)$</td>
</tr>
<tr>
<td>filter</td>
<td>a</td>
<td>$x_{vis}, x_{tst}$</td>
<td>att</td>
<td>$a_{out} = \text{and}(a, \text{find}[x_{vis}, x_{tst}])$, i.e. reusing find and and</td>
</tr>
<tr>
<td>[exist, count]</td>
<td>a</td>
<td>(none)</td>
<td>ans</td>
<td>$y = W^T \text{vec}(a)$</td>
</tr>
<tr>
<td>describe</td>
<td>a</td>
<td>$x_{vis}, x_{tst}$</td>
<td>ans</td>
<td>$y = W_2^T (W_2 \text{sum}(a \odot x_{vis}) \odot W_3 x_{tst})$</td>
</tr>
<tr>
<td>[eq.count, more, less]</td>
<td>$a_1, a_2$</td>
<td>(none)</td>
<td>ans</td>
<td>$y = W_2^T \text{vec}(a_1) + W_2^T \text{vec}(a_2)$</td>
</tr>
<tr>
<td>compare</td>
<td>$a_1, a_2$</td>
<td>$x_{vis}, x_{tst}$</td>
<td>ans</td>
<td>$y = W_1^T (W_2 \text{sum}(a_1 \odot x_{vis}) \odot W_3 \text{sum}(a_2 \odot x_{vis}) \odot W_4 x_{tst})$</td>
</tr>
</tbody>
</table>

• Modules are dynamically assembled into networks on-the-fly

"How many other things are of the same size as the green matte ball?"
Learning from Expert Layouts

**Stage 1:** train the model to predict the ground-truth (gold) layout with supervised learning (behavioral cloning from expert layouts)

*is the material of the yellow block same as the yellow cylinder?*

**Expert (Gold) Layout:** [Find, Find, Compare]

**Module 0:** Find yellow cylinder

**Module 1:** Find yellow block

**Module 2:** Compare material

**no**

Expert (gold) layout from
- dataset annotations or
- syntactic parsing
End-to-End Layout Search with Policy Gradients

Stage 2: sample multiple candidate layouts from the layout policy, and optimize with policy gradient (REINFORCE)

- Sampled Layout 1: [Find, Find, Compare]
- Sampled Layout 2: [Find, Describe]
- Sampled Layout 3: [Find, Transform, Describe]

Ground-truth answer: “no”

<table>
<thead>
<tr>
<th>Sampled Layout</th>
<th>Prediction</th>
<th>QA Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Find, Find, Compare]</td>
<td>“no”</td>
<td>0.05</td>
</tr>
<tr>
<td>[Find, Describe]</td>
<td>“yes”</td>
<td>1.72</td>
</tr>
<tr>
<td>[Find, Transform, Describe]</td>
<td>“no”</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Qualitative results on the CLEVR dataset (synthetic images)

Does the blue cylinder have the same material as the big block on the right side of the red metallic thing?
Qualitative results on the CLEVR dataset (synthetic images)

question: *do the small cylinder that is in front of the small green thing and the object right of the green cylinder have the same material?*

ground-truth answer: *no*

**Stage 1**
clone expert (gold) layout

**Stage 2**
end-to-end layout search
Quantitative results on the CLEVR dataset (synthetic images)

- Superior performance with end-to-end training
Qualitative results on the VQA dataset (natural images)

What is on the table?
Qualitative results on the VQA dataset (natural images)

What color is the plate?
Qualitative results on the VQA dataset (natural images)

What is behind the foot of the bed?
Quantitative results on the VQA dataset (natural images)

• Works well on real images and questions
Summary of N2NMN

- Discrete compositionality with trainable and reusable modules
- Jointly train policy (*what*) and compositional modules (*how*)
- A possible way to bridge neural + symbolic

```python
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```

What color is the thing with the same size as the blue cylinder?

Output answer: “gray”
Fine-grained Textual Explanations

Trevor Darrell
UC Berkeley

With Lisa Anne Hendricks, Zeynep Akata, Ronghang Hu, Bernt Schiele, Marcus Rohrbach, …
Cardinal
What type of bird is this?

It is a **Cardinal** because it is a **red bird** with a **red beak** and a **black face**.
This is a **White Necked Raven** because it is a black bird with a white nape and a large beak.

Hendricks et al. Generating Counterfactual Explanations with Natural Language. ICML Workshops 2018.
Fine-grained Explanations
Fine-grained Explanations

Fine-grained Classification Model

Deep Model

Label: Cardinal
Fine-grained Explanations

Finegrained Classification Model → Cardinal → LSTM
This is a *Cardinal* because this bird has a red crown, a short bill, and a red breast.

Descriptions from: Reed et al. Learning deep representations of fine-grained visual descriptions. CVPR 2016.
This is a *Cardinal* because this bird has a red crown, a short bill, and a red breast.
This is a *Cardinal* because this bird has a red crown, a short bill, and a red breast.
What makes a good visual explanation?

This is a *Cardinal* because this bird has a red crown, a short bill, and a red breast.
What makes a good visual explanation?

This is a *Cardinal* because this is a red bird with a black face and a red beak.
Visual Explanation Model

Finegrained Classification Model

Cardinal

LSTM
Visual Explanation Model

Finegrained Classification Model

Cardinal

LSTM
Target sentence:
This is a red bird with a black cheek patch.
Target sentence:
This is a red bird with a black cheek patch.
Text Discriminator Model

...because this is a red bird with a black face and a red beak.

Cardinal
Target sentence:
This is a red bird with a black cheek patch.
Target sentence:
This is a red bird with a black cheek patch.

This is a **White Necked Raven**...

*Description:* because this bird is nearly all black with a short pointy bill.
This is a **White Necked Raven**...

*Description*: because this bird is nearly all black with a short pointy bill.

*Explanation*: because this is a black bird with a white nape and a large black beak.
Choose the image which most closely matches the following text:

… this is a black bird with a white nape and a large black beak.
Choose the image which most closely matches the following text:

… this is a black bird with a white nape and a large black beak.
Which model is best for discriminating between images?

![Bar chart comparing descriptions and explanations with % Correctly Selected Image on the x-axis.](chart.jpg)

Higher is better.
Which of the following is the best explanation for why this bird is a White Necked Raven?

A) This is a White Necked Raven because this bird is nearly all black with a short pointy bill.
B) This is a White Necked Raven because this is a black bird with a white nape and a large black beak.
Which of the following is the best explanation for why this bird is a White Necked Raven?

A) This is a **White Necked Raven** because this bird is nearly all black with a short pointy bill.

B) This is a **White Necked Raven** because this is a black bird with a white nape and a large black beak.
Which of the following is the best explanation for why this bird is a White Necked Raven?

A) This is a **White Necked Raven** because this bird is nearly all black with a short pointy bill.
B) This is a **White Necked Raven** because this is a black bird with a white nape and a large black beak.

Need bird watchers!
Which explanations do bird watchers prefer?
This is a mallard because this is a brown and white bird with a green head and a yellow beak.
This is a *mallard* because this is a brown and white bird with a green head and a yellow bill.
This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.
Intuition: Only output explanations which are grounded in visual evidence.
This is a **mallard** because this is a *brown and white bird* with a green head and a *yellow bill*.
This is a mallard because this is a brown and white bird with a green head and a yellow bill.
This is a **mallard** because this is a **brown and white bird** with a **green head** and a **yellow bill**.

- Call “brown and white bird”, “green head”, and “yellow bill” attributes.
- Extract attributes with a noun phrase chunker.
Query: A brightly colored umbrella.

Model from:
Query: A brightly colored umbrella.

Model from:

Trained with Visual Genome:
Explanation Sampler

Finegrained Classification Model

Mallard

LSTM
This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.
Generally score sentences based off *sentence fluency*:

\[ S = \sum_{t} \log P(w_t|w_{0:t-1}, I, C) \]
Generally score sentences based off sentence fluency:

$$S = \sum_{t} \log P(w_t | w_{0:t-1}, I, C)$$
Can we score sentences on visual grounding instead?

This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.
This is a mallard because this bird has a brown head, orange feet, and a flat bill.

Score for brown head: 1.9  
Score for orange feet: 2.1  
Score for flat bill: 1.1  

Average score high → good explanation.
This is a mallard because this is a brown and white bird with a green head and a yellow bill.

Score for brown and white bird: 2.2
Score for green head: 0.2
Score for yellow beak: 1.2

Average score low → bad explanation.
Score sampled sentences with visual grounding model.

$$S = \frac{1}{|A|} \sum_{a \in A} GroundingScore (a, I)$$

A is set of attributes in explanation

This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.
This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.

A is set of attributes in explanation

$$S = \frac{1}{|A|} \sum_{a \in A} GroundingScore(a, I)$$

Grounding score for attribute in image.
This is a **Eared Grebe** because …. 

**Baseline:**
this is a black bird
with a long neck and
red eyes.

**Average Grounding:**
…this is a **black bird**
with a **white eye** and a
**red eye**.
This bird has a brown head, orange feet, and a flat bill.

Grounding Model

Brown head  (score: 1.2)
Flat bill     (score: 0.8)
Orange feet  (score: 0.9)
This bird has a brown head, orange feet, and a flat bill.

Grounding Model

Brown head (score: 1.2)
Flat bill (score: 0.8)
Orange feet (score: 0.9)

Phrase Critic

LSTM
Brown head S: 1.2
LSTM
Flat bill S: 0.8
LSTM
Orange feet S: 0.9

Score: 2.05
Positive sentence: This bird has a brown head, orange feet, and a flat bill.

Negative sentence: This bird has a brown head, black feet, and a flat bill.
Positive sentence: This bird has a brown head, orange feet, and a flat bill.

Negative sentence: This bird has a brown head, black feet, and a flat bill.

Grounding Model

Phrase Critic
Score: 2.05

Phrase Critic
Score: 1.02

Ranking Loss
Positive Sentence: This bird has a brown head, orange feet and a flat bill.
Positive Sentence: This bird has a brown head, orange feet and a flat bill.
Positive Sentence: This bird has a brown head, orange feet and a flat bill.

Negative Sentence: This bird has a brown head, black feet and a flat bill.
Score sampled sentences with phrase critic.

\[ S = PhraseCritic(A, I) \]
Extracted noun phrase from explanation: brown and white bird, green head, yellow bill.
Does this bird have a **green head**?
Does this bird have a green head? ✗
Are grounded explanations more image relevant?

Higher is better.

- Baseline
- Average Grounding
- Phrase Critic

% Sentences with Correct Noun Phrases
This is a **Eared Grebe** because ….

**Baseline:**
this is a black bird with a long neck and red eyes

**Average grounding:**
this is a **black bird** with a **white eye** and a **red eye**.

**Phrase critic:**
this bird has a **long neck** and **bright orange eyes**.
This is a **Evening Grosbeak** because ….

*Baseline:* this is a yellow bird with a black and white wing and a yellow beak.

*Average grounding:* this is a **white bird** with a **brown and black wing** and a **yellow beak**.

*Phrase critic:* this is a small brown bird with a white and black wing and a **yellow beak**.
What type of bird is this?

It is a Cardinal because it is a red bird with a red beak and a black face.
Why isn’t it a Scarlet Tanager?

It isn’t a *Scarlet Tanager* because it doesn’t have black wings.
Pipeline:

Why isn’t this a Scarlet Tanager?
Pipeline:

Why isn’t this a Scarlet Tanager?

Predict evidence for Scarlet Tanager:

This is a red bird with black wings. This red bird has a pointy beak and black eyes.

...
Pipeline:

Why isn’t this a Scarlet Tanager?

Predict evidence for Scarlet Tanager:

This is a red bird with black wings. This red bird has a pointy beak and black eyes. ...

Ground Scarlet Tanager evidence:

Red bird: grounded
Pointy beak: grounded
...
Black wings: Not grounded!
Pipeline:

Why isn’t this a Scarlet Tanager?

Predict evidence for Scarlet Tanager:

This is a red bird with black wings. This red bird has a pointy beak and black eyes.

Ground Scarlet Tanager evidence:

Red bird: grounded
Pointy beak: grounded
... Black wings: Not grounded!

Construct sentence:

This is not a Scarlet Tanager because it does not have black wings.
Why is this a *Blue Winged Warbler* and not a *Common Yellowthroat*?

Explanation: This is a *Blue Winged Warbler* because this is a yellow bird with a black wing and a black pointy beak.

This is not a *Common Yellowthroat* because it does not have a black face.
Are Explanations Helpful to Humans?
Are Explanations Helpful to Humans?

The AI justified its prediction with the following evidence: this is a brown and black spotted bird with a white belly. Do you think you would accept the AI’s prediction?

- Accept prediction
- Do not accept prediction
Are Explanations Helpful to Humans?

The AI is **wrong**; you **should not** accept the prediction.

The AI justified its prediction with the following evidence: this is a **brown and black spotted bird** with a **white belly**. Do you think you would accept the AI's prediction?

- [ ] Accept prediction
- [x] Do not accept prediction
Are Explanations Helpful to Humans?

The AI justified its prediction with the following evidence: this is a small orange bird with a black wing and a small black beak. Do you think you would accept the AI's prediction?

- Accept prediction
- Do not accept prediction
Are Explanations Helpful to Humans?

The AI is *correct*; you *should* accept the prediction.

The AI justified its prediction with the following evidence: this is a *small orange bird* with a *black wing* and a *small black beak*. Do you think you would accept the AI's prediction?

- Accept prediction
- Do not accept prediction
Are Explanations Helpful to Humans?

Correctly Accepted/Rejected Decision

<table>
<thead>
<tr>
<th>Explanation</th>
<th>45</th>
<th>50</th>
<th>55</th>
<th>60</th>
<th>65</th>
</tr>
</thead>
<tbody>
<tr>
<td>No explanation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Are Explanations Helpful to Humans?

Correctly Accepted/Rejected Decision

No explanation

Explanation

45  50  55  60  65
Are Explanations Helpful to Humans?

Correctly Accepted/Rejected Decision
What makes a good visual explanation?
What makes a good visual explanation?
What makes a good visual explanation?

- Grounding
- Discriminative Loss
- Image Relevance
- Class Relevance
- Description
- Visual Explanation
- Definition
Driving-X

Image credit: Berkeley Deep Drive

Image credit: H. Miller, 1957

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata

UC Berkeley
University of Amsterdam
eXplainable AI (for self-driving cars)

Need **introspective** or **debuggable** driving model:

Explanations are grounded in the network’s true internal state.

**Why?**
1) Requires a very high level of **trust**.
2) Users should be able to **anticipate** what the vehicle will do.
3) Effective human-machine communication.
Outline

- **Interpretable Learning for Self-driving Cars by Visualizing Causal Attention**

- **Textual Explanations for Self-driving Vehicles**
  Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata, ECCV 2018.

- **Internalizing Human-to-Vehicle Advice for Self-driving Vehicles**

- **Advisable Learning for Self-driving Vehicles**
  Jinkyu Kim, Anna Rohrbach, Dequan Wang, Trevor Darrell, and John Canny, *under review.*
Visualizing Causal Attention
Visualizing Causal Attention

Highlights image regions that causally influence the network’s output (i.e., steering)

[Kim and Canny, ICCV’17]
Fine-Grained Decoder (Causality check)

Fine-grained decoder to remove spurious attention blobs and to find causal local visual blobs

[Kim and Canny, ICCV'17]
Examples of Attention Map

Attention maps over time (from left to right)

[Kim and Canny, ICCV'17]
Control accuracy is not degraded by incorporation of attention compared to an identical base CNN without attention.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>MAE in degree [SD]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Testing</td>
</tr>
<tr>
<td>Comma.ai</td>
<td>CNN+FCN</td>
<td>.421 [0.82]</td>
</tr>
<tr>
<td></td>
<td>CNN+LSTM</td>
<td>.488 [1.29]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=0)</td>
<td>.497 [1.32]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=10)</td>
<td>.464 [1.29]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=20)</td>
<td>.463 [1.24]</td>
</tr>
<tr>
<td>HCE</td>
<td>CNN+FCN</td>
<td>.246 [.400]</td>
</tr>
<tr>
<td></td>
<td>CNN+LSTM</td>
<td>.568 [.977]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=0)</td>
<td>.334 [.766]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=10)</td>
<td>.358 [.728]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=20)</td>
<td>.373 [.724]</td>
</tr>
<tr>
<td>Udacity</td>
<td>CNN+FCN</td>
<td>.457 [.870]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=0)</td>
<td>.491 [1.20]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=10)</td>
<td>.489 [1.19]</td>
</tr>
<tr>
<td></td>
<td>Attention (λ=20)</td>
<td>.489 [1.26]</td>
</tr>
</tbody>
</table>

[Kim and Canny, ICCV'17]
Causal Attention Heat Maps

- Raw input image
- Visual attention heatmaps
  - with spurious attention sources
- Attention heat maps by filtering out spurious blobs

[Kim and Canny, ICCV'17]
Textual Explanations

Can I park here?

No!

Why?

Because there is a red curb, which indicates no parking!
Textual Explanations

Example of textual descriptions + explanations:

**Ours:** “The car is driving forward + because there are no other cars in its lane”

**Human annotator:** “The car heads down the street + because the street is clear.”

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV’18]
Berkeley DeepDrive Video (BDD-V) Data

Over 10,000 hours of driving data, which provides (1) dash-cam video, (2) GPS, (3) course and speed

[Xu, Gao, Yu, Darrell, CVPR'17]
Berkeley DeepDrive eXplanation (BDD-X) dataset

**A**

Input images

1. The car is driving
2. The car is moving into the right lane
3. The car moves back into the left lane
4. The car drives in the left lane
5. The car moves into the right lane

**Action descriptions:**

- (1) The car is driving
- (2) The car is moving into the right lane
- (3) The car moves back into the left lane
- (4) The car drives in the left lane
- (5) The car moves into the right lane

**Action explanations:**

- as there is nothing to impede it.
- because it is safe to do so.
- because the school bus in front of it is stopping.
- in order to pass the school bus.
- since it has now passed the school bus and it is taking the right fork.

**B**

<table>
<thead>
<tr>
<th>DRV-X dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Frames</td>
<td>8,400,000</td>
</tr>
<tr>
<td>Hours</td>
<td>≈ 77 hours</td>
</tr>
<tr>
<td>Condition</td>
<td>Urban</td>
</tr>
<tr>
<td>Lighting</td>
<td>Day/Night</td>
</tr>
<tr>
<td># Annotations</td>
<td>26,228</td>
</tr>
<tr>
<td>Avg. # actions/videos</td>
<td>3.8</td>
</tr>
<tr>
<td># Videos</td>
<td>6,984</td>
</tr>
<tr>
<td># Training</td>
<td>5,588</td>
</tr>
<tr>
<td># Validation/Testing</td>
<td>696</td>
</tr>
</tbody>
</table>

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
Berkeley DeepDrive eXplanation (BDD-X) dataset

<table>
<thead>
<tr>
<th>BDD-X action descriptions</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop</td>
<td>6879</td>
</tr>
<tr>
<td>slow</td>
<td>6122</td>
</tr>
<tr>
<td>forward</td>
<td>4322</td>
</tr>
<tr>
<td>drive</td>
<td>3994</td>
</tr>
<tr>
<td>move</td>
<td>3273</td>
</tr>
<tr>
<td>accelerate</td>
<td>2882</td>
</tr>
<tr>
<td>right</td>
<td>2616</td>
</tr>
<tr>
<td>left</td>
<td>2574</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BDD-X action explanations</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>traffic</td>
<td>7486</td>
</tr>
<tr>
<td>light</td>
<td>6116</td>
</tr>
<tr>
<td>red</td>
<td>3979</td>
</tr>
<tr>
<td>move</td>
<td>3915</td>
</tr>
<tr>
<td>clear</td>
<td>3660</td>
</tr>
<tr>
<td>ahead</td>
<td>3629</td>
</tr>
<tr>
<td>road</td>
<td>3528</td>
</tr>
<tr>
<td>stop</td>
<td>3430</td>
</tr>
</tbody>
</table>
Two approaches (SAA and WAA) to align the vehicle controller and the textual justifier such that they look at the same input regions.

**Model**

"The car slows down because it's turning to the right."

**SAA: Use vehicle controller's attention**

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
Model

Two approaches (SAA and WAA) to align the vehicle controller and the textual justifier such that they look at the same input regions.

**Vehicle controller**

- Spatial attention
- CNN
- Attention alignment loss

**SAA**

**WAA**

- Explanation generator has its own attention, which is tied with vehicle controller's attention via a loss function

“The car is slowing down + because there is a stop sign ahead to turn.”

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
Quantitative Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda_c$</th>
<th>Mean of absolute error (MAE)</th>
<th>Mean of distance correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acceleration (m/s$^2$)</td>
<td>Course (degree)</td>
</tr>
<tr>
<td>CNN+FC [1]$^\dagger$</td>
<td>-</td>
<td>6.92 [7.50]</td>
<td>12.1 [19.7]</td>
</tr>
<tr>
<td>CNN+FC [1]+P</td>
<td>-</td>
<td>6.09 [7.73]</td>
<td>6.74 [14.9]</td>
</tr>
<tr>
<td>CNN+LSTM+Attention [4]$^\dagger$</td>
<td>-</td>
<td>6.87 [7.44]</td>
<td>10.2 [18.4]</td>
</tr>
<tr>
<td>CNN+LSTM+Attention+P (Ours)</td>
<td>1000</td>
<td>5.02 [6.32]</td>
<td>6.94 [15.4]</td>
</tr>
<tr>
<td>CNN+LSTM+Attention+P (Ours)</td>
<td>100</td>
<td>2.68 [3.73]</td>
<td>6.17 [14.7]</td>
</tr>
<tr>
<td>CNN+LSTM+Attention+P (Ours)</td>
<td>10</td>
<td>2.33 [3.38]</td>
<td>6.10 [14.7]</td>
</tr>
<tr>
<td>CNN+LSTM+Attention+P (Ours)</td>
<td>0</td>
<td>2.29 [3.33]</td>
<td>6.06 [14.7]</td>
</tr>
</tbody>
</table>

- Prior measurements (P) help
- Spatial attention helps
- Low entropy attention leads to higher error

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
## Quantitative Analysis

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Control inputs</th>
<th>Explanations (e.g. “because the light is red”)</th>
<th>Descriptions (e.g. “the car stops”)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\lambda_s$</td>
<td>$\lambda_c$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-XAI baseline</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2VT [17]</td>
<td>N</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2VT +SA</td>
<td>N</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2VT +SA +TA</td>
<td>N</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Rationalization</td>
<td>Ours (no constraints)</td>
<td>Y</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Ours (with SAA)</td>
<td>Y</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Ours (with SAA)</td>
<td>Y</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Ours (with SAA)</td>
<td>Y</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Intraspective</td>
<td>Ours (with WAA)</td>
<td>Y</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>explanation</td>
<td>Ours (with WAA)</td>
<td>Y</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Ours (with WAA)</td>
<td>Y</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

- All better than baseline
- Intraspective better than Rationalization
- WAA is best

**Abbreviation:**
- S2VT (seq-to-seq video-to-text)
- TA (temporal fusion)
- SA (spatial attention)
- WAA (Weakly aligned attention)
- SAA (Strongly aligned attention)

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
## Human Evaluation

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Control inputs</th>
<th>$\lambda_a$</th>
<th>$\lambda_c$</th>
<th>Correctness rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Explanations</td>
</tr>
<tr>
<td><strong>Non-XAI baseline$^\dagger$</strong></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22.4%</td>
</tr>
<tr>
<td><strong>Rationalization</strong></td>
<td>Ours (no constraints)</td>
<td>Y</td>
<td>0</td>
<td>0</td>
<td>64.0%</td>
</tr>
<tr>
<td><strong>Introspective explanation</strong></td>
<td>Ours (with SAA)</td>
<td>Y</td>
<td>-</td>
<td>100</td>
<td>62.4%</td>
</tr>
<tr>
<td></td>
<td>Ours (with WAA)</td>
<td>Y</td>
<td>10</td>
<td>100</td>
<td><strong>66.0%</strong></td>
</tr>
</tbody>
</table>

Table 3: Human evaluation of the generated action descriptions and explanations for randomly chosen 250 video intervals. We measure the success rate where at least 2 human judges rate the generated description or explanation with a score 1 (correct and specific/detailed) or 2 (correct). $^\dagger$: Sentences are sampled based on their frequency in the training data (i.e. a strong prior).
Examples of explanations generated

Human: The car steadily driving → now that the cars are moving.
Ours (WAA): The car is driving forward → because traffic is moving freely.
Ours (SAA): The car heads down the road → because traffic is moving at a steady pace.
Rationalization: The car slows down → because it’s getting ready to a stop sign.

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
Examples of explanations generated

**Human:** The car slows down + since it is about to turn left.

**Ours (WAA):** The car slows down + because it is preparing to turn to the road.

**Ours (SAA):** The car is slowing + because it is approaching a stop sign.

**Rationalization:** The car slows + because there is a stop sign.

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
Examples of explanations generated

Human: The car slows down + since it is about to turn left.

Ours (WAA): The car slows down + because it is preparing to turn to the road.

Ours (SAA): The car is slowing + because it is approaching a stop sign.

Rationalization: The car slows + because there is a stop sign.

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
Examples of explanations generated

**Human:** The car is stopped because it waits for traffic in front of it to move.

**Ours (WAA):** The car is stopped because traffic is stopped.

**Ours (SAA):** The car is stopped because the car in front of it is stopped.

**Rationalization:** The car is stopped because it is parked in the left lane.

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]
eXplainable AI (for self-driving cars)

- Visualizing **Attention** Maps
- Generating Textual **Explanations**
Advisable AI (for self-driving cars)

We want to allow end-users to not only understand the controller, but to influence it.
Why Advice and Not Commands?

- Users will not be aware of the full state of the vehicle (they are not driving), controller should be in charge.

- Users have real world knowledge that the controller lacks which can improve safety and ride quality.
Why Advice and Not Commands?

- Advice will often be given offline, or at the beginning of a ride:
  - When at an intersection, look out for pedestrians
  - Drive gently (occupant gets carsick)

- What we have now: instantaneous control using text grounded in the video.
Advisable AI (for self-driving cars)
Advisable AI (for self-driving cars)
Qualitative results

<table>
<thead>
<tr>
<th>Input image</th>
<th>Attention heat maps</th>
</tr>
</thead>
</table>
|             | Baseline [6]  
(without taking advice) | Ours  
(trained with $ADV_S$) | Ours  
(trained with $ADV_C + ADV_S +$<none>) |
| ![Image](image1.png) | ![Heatmap](heatmap1.png) | ![Heatmap](heatmap2.png) | ![Heatmap](heatmap3.png) |
| -2.1° | 38.14 km/h | -1.9° | 31.9 km/h | -2.8° | 33.8 km/h | -2.8° | 33.8 km/h |

: “There is a white car pulling out of a parking lot.”

<table>
<thead>
<tr>
<th>Input image</th>
<th>Attention heat maps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td>3.8°</td>
<td>25.3 km/h</td>
</tr>
</tbody>
</table>

: “After the turn, be cautious due to pedestrian crosswalks.”
## Qualitative results

### Input image

<table>
<thead>
<tr>
<th>Advice</th>
<th>Steering angle</th>
<th>Speed</th>
<th>Input image</th>
<th>Attention heat maps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wait at a pedestrian crosswalk while a <strong>pedestrian crossing</strong> and then drive forward and change to the left lane.</td>
<td>2.9°</td>
<td>24.36 km/h</td>
<td><img src="image1" alt="Input image" /></td>
<td><img src="heatmaps1" alt="Attention heat maps" /></td>
</tr>
<tr>
<td>See <strong>cones</strong> to the left of the turn lane.</td>
<td>-3.8°</td>
<td>14.09 km/h</td>
<td><img src="image2" alt="Input image" /></td>
<td><img src="heatmaps2" alt="Attention heat maps" /></td>
</tr>
</tbody>
</table>

### Baseline [6] (without taking advice)

<table>
<thead>
<tr>
<th>Speed</th>
<th>Steering angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.4 km/h</td>
<td>-4.3°</td>
</tr>
</tbody>
</table>

### Ours (trained with ADVg)

<table>
<thead>
<tr>
<th>Speed</th>
<th>Steering angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9 km/h</td>
<td>-1.8°</td>
</tr>
</tbody>
</table>

### Ours (trained with ADVg + ADVg + <none>)

<table>
<thead>
<tr>
<th>Speed</th>
<th>Steering angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.7 km/h</td>
<td>6.8°</td>
</tr>
</tbody>
</table>
Fine-grained Attention / Long-term Advice

We have explored an explainable and advisable driving model, which we explore several approaches for better forms of explainability:

- We use semantic segmentation as an input representation.
- We update model to take long-term advice (i.e offline driving instructions).

Instance Attention

We are exploring a composite end-to-end vehicle controller that integrates our advisable/explainable model with our recent object-centric attention model (ICRA '19).

Initial experimental result
Today

• Multi-step Saliency via Compositional NMNs
• Fine-grained Textual Explanations
• From Explainable to “Advisable” Driving Models