Programming your way to explainable AI

Mark Hammond
Co-founder / CEO
“No one really knows how the most advanced algorithms do what they do. That could be a problem.”

Will Knight
*The Dark Secret at the Heart of AI*

https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/
Expert systems artificial intelligence vs. modern machine learning

Programming your way to explainable AI
Why do we want to do this?

• Users already expect to be able to audit systems (including human systems)
• The errors ML algorithms make are very “foreign” to many users
• Understanding why an error occurred is the first step to fixing it (debugging)
• Trust, Security, and Confidence
• Verification of behavior
• Debug and Refine
• Prevent Biases
Okay, we’ve established what and why, but to what extent?

- Constant refrain of “why?”
- What level of abstraction?
- Introspection vs. Justification

- Do we want the systems we build to be able to explain the features they've learned and how they were applied?
- Do we want the system to justify why a prediction was reasonable?
- Do we want both?
- How can we work towards achieving these objectives?

[Link](http://www.scq.ubc.ca/a-dialogue-with-sarah-aged-3-in-which-it-is-shown-that-if-your-dad-is-a-chemistry-professor-asking-%e2%80%9cwhy%e2%80%9d-can-be-dangerous-5/)
Deep Explanations

• Try to tease apart what is happening within the neural network itself

• Learning Semantic Associations
• Generating Visual Explanations
• Rationalizing Neural Predictions
Learning Semantic Associations

**Goal:** “Count” occurrences of items of interest in frames of videos so that we can later search video content.

Learning Semantic Associations

Concepts

- Visual
  - object, scene, action and activity concepts
- Audible
- ASR/OCR Text

Stack Convolutional layers for classification.

Multimedia Event Recounting

- This illustrates and example of event recounting.
- The system classified this video as a wedding.
- The frames above show its evidence for the wedding classification

Learning Semantic Associations

- Train the net to associate semantic attributes with hidden layer nodes
- Train the net to associate labelled nodes with known ontologies
- Generate examples of prominent but unlabeled nodes to discover semantic labels
- Generate clusters of examples from prominent nodes
- Identify the best architectures, parameters, and training sequences to learn the most interpretable models
Generating Visual Explanations

**Goal:** Generate a textual description (a caption) about why an image is in a class

- *Introspection versus Justification*
- *Class discriminative and accurate*

[Diagram showing images and descriptions of Western Grebe and Laysan Albatross]

- **Western Grebe**
  - **Description:** This is a large bird with a white neck and a black back in the water.
  - **Class Definition:** The *Western Grebe* is a waterbird with a yellow pointy beak, white neck and belly, and black back.
  - **Explanation:** This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

- **Laysan Albatross**
  - **Description:** This is a large flying bird with black wings and a white belly.
  - **Class Definition:** The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
  - **Visual Explanation:** This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.
  - **Description:** This is a bird with a white neck and a black back in the water.
  - **Class Definition:** The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
  - **Visual Explanation:** This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.
Generating Visual Explanations

- Concatenate the CNN classification of the image with the target sentence and image category
- Feed to LSTMs to generate description
- Reward the model on how much a statement discriminates between other classes
- And on Relevance (the probability the selected words are correct)
Rationalizing Neural Predictions

Goal: extract a snippet of phrases or words that explain a classification

<table>
<thead>
<tr>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>the beer was n't what i expected, and i'm not sure it's “true to style”, but i thought it was delicious. a very pleasant ruby red-amber color with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Look: 5 stars</td>
</tr>
</tbody>
</table>

what is the easiest way to install all the media codec available for ubuntu ? i am having issues with multiple applications prompting me to install codecs before they can play my files. how do i install media codecs ?

what should i do when i see <unk> report this <unk> ? an unresolvable problem occurred while initializing the package information. please report this bug against the 'update-manager' package and include the following error message: e : encountered a section with no package : header e : problem with mergelist <unk> : the package lists or status file could not be parsed or opened.

please anyone give the solution for this whenever i try to convert the rpm file to deb file i always get this problem error : <unk> : not an rpm package ( or package manifest ) error executing " lang="rpm -qa --queryformat "% ( name ) <unk> "" : at <unk> line 489 thanks converting rpm file to debian file.

how do i mount a hibernated partition with windows 8 in ubuntu ? i ca n't mount my other partition with windows 8 , i have ubuntu 12.10 amd64 : error mounting /dev/sda1 at <unk> : command-line 'mount -t ntfs -o ' uhelper=udisks2 , nodev , nosuid , uid=1000 , gid=1000 , dmask=01777 , fmask=01777 " /dev/sda1 " <unk> " exited with non-zero exit status 14 : windows is hibernated , refused to mount . failed to mount '/dev/sda1' : operation not permitted the ntfs partition is hibernated . please resume and shutdown windows properly , or mount the volume read-only with the 'ro' mount option.

Goal: highlight the most pertinent part of questions of the Ubuntu user forums

A classifier (in this case an RNN for text classification)

I really did not this. It just seemed extremely watery. I don’t think this had any carbonation whatsoever. Maybe it was flat, who knows? But even if I got a bad brew I don’t see how this would possibly be something I’d get time and time again. I could taste the hops towards the middle, but the beer got pretty nasty towards the bottom. I would never drink this again, unless it was free. I’m kind of upset I bought this.

Extract text with two goals:
1. Continuity
2. Brief and coherent (implemented as RCNN)
Model Induction

• Look at the behavior of the resulting trained system to infer a model that can be used to explain the observed behavior

• Local Interpretable Model-Agnostic Explanations (LIME)
• Anchor Local Interpretable Model-Agnostic Explanations (aLIME)
Local Interpretable Model-Aagnostic Explanations (LIME)

**Goal:** Highlight salient characteristics in input data that lead to a given classification

Two levels of trust
- Trust the prediction: LIME
- Trust the model: SP-LIME

Figure 4: Explaining an image classification prediction made by Google’s Inception network, highlighting positive pixels. The top 3 classes predicted are “Electric Guitar” ($p = 0.32$), “Acoustic guitar” ($p = 0.24$) and “Labrador” ($p = 0.21$)

Local Explainability

- Select subsets of input (weighted random sample) and classify them
- Of possible subsets optimize for explainability
  - The faithfulness of the classification within the subset of the data
  - The interpretability by humans
- Requires different approach for different data types.
  - Text: bag of words
  - Image: bag of super pixels
- Intelligently select multiple local explanations for global explainability
Anchor Local Interpretable Model-Agnostic Explanations (aLIME)

- Create local explanations using if/then rules that are “anchored” such that changes to the rest of the instance do not matter.

![Diagram of LIME explanation](https://arxiv.org/abs/1611.05817)

Figure 1: Explaining a prediction from the UCI adult dataset. The task is to predict if a person’s salary is higher than 50,000 dollars (>50k) or not (<50K).

https://arxiv.org/abs/1611.05817
Anchor Local Interpretable Model-Agnostic Explanations (aLIME)

(a) Original image  
(b) Anchor for “Zebra”  
(c) Images with $P(\text{zebra}) > 90\%$

Figure 5: **Image classification**: explaining a prediction from Inception, and examples from $\mathcal{D}(z|c, x)$
Integrating human intelligence and machine intelligence
Recomposability and behaviors

Simulation or Physical System

Actor -> Action -> Environment -> Assessment

State

Reward
Recomposability and behaviors

- Re-imagine the subsumption architecture (Rodney Brooks, 1986) with deep reinforcement learning
- Decompose behaviors into sub-behaviors and organize into a hierarchy of layers
- Kind of like semantic associations--each layer can explain itself
Concept graphs and Inkling

• A concept can either represent a skill or a feature of the world

• Each concept has its own training regime

• Skill concepts are composed through selection or synthesis

• Explainability means expressing how a skill contributed to the result and how much it contributed

[Diagram of a concept graph]

Programming your way to explainable AI
What does this look like
Bringing it all together

• Rapid iteration in this space
• Deep Explanations
• Model Induction
• Machine Teaching
• Recomposability
Join us for the Journey

• Do you have a control or optimization problem you’re looking to tackle with a need for explainability?

  Sign up for our recently announced early adopter program https://bons.ai/getting-started

• Are you interested in working in this area?

  We’re hiring!
  https://bons.ai/careers

• Questions? Want to learn more?

  Visit us at our booth