Natural Language Explanations of Deep Networks

Jacob Andreas
MIT CSAIL
LINGO.csail.mit.edu
Deep learning is hard
Deep learning is hard

What is this a picture of?

[Photo credit: Lucy Li (Wikipedia), dialogue: http://demo.visualdialog.org/]
Deep learning is hard

What is this a picture of?

it looks like a <unknown_word> of some sort

[Photo credit: Lucy Li (Wikipedia), dialogue: http://demo.visualdialog.org/]
Deep learning is hard

What is this a picture of?

It looks like a <unknown_word> of some sort

How many giraffes are in this picture?

[Photo credit: Lucy Li (Wikipedia), dialogue: http://demo.visualdialog.org/, giraffe question: Janelle Shane]
Deep learning is hard

What is this a picture of?

it looks like a <unknown_word> of some sort

How many giraffes are in this picture?

too many to count

[Photo credit: Lucy Li (Wikipedia), dialogue: http://demo.visualdialog.org/, giraffe question: Janelle Shane]
Explaining deep networks is hard
Explaining deep networks is hard

electric guitar  [Ribeiro et al. 2016. LIME]
Explaining deep networks is hard

electric guitar

[Ribeiro et al. 2016. LIME]
Explaining deep networks is hard.

[Olah et al. 2017. The building blocks of interpretability.]
Explaining deep networks is hard

[Olah et al. 2017. The building blocks of interpretability.]

[electric guitar] [Ribeiro et al. 2016. LIME]
Explaining deep networks is hard

What aspect of this region is relevant?

What do these pictures have in common?

electric guitar [Ribeiro et al. 2016. LIME.]

[Olah et al. 2017. The building blocks of interpretability.]
Explaining deep networks is hard

moderate pulmonary edema

What are explanations for?

Communicating:
Model safety and potential vulnerabilities
Changes needed to produce desired outputs
Actions that *users* should take in collaborative settings
Communicating explanations

the fretboard is gray
mammals with black noses
interstitial abnormality with peribronchial cuffing

Explanation as translation
Multi-agent communication
[Sukhbaatar et al. 16. Learning multiagent communication with backpropagation.]
Translating neuralese

all clear
A statistical MT problem

\[
\max p(\theta | a) \cdot p(a)
\]

[1.0 2.3]  
[-0.3 0.4]  
[-1.2 1.1]

all clear

[e.g. Koehn 10. Phrase-based statistical machine translation.]
A statistical MT problem

\[
\max p(\emptyset | a) p(\{a\})
\]

"looks like a translation"

\[
\begin{array}{cc}
1.0 & 2.3 \\
-0.3 & 0.4 \\
-1.2 & 1.1 \\
\end{array}
\]

"looks like English"

all clear

[e.g. Koehn 10. Phrase-based statistical machine translation.]
A statistical MT problem

$p(\text{a} | \text{a})$

\begin{align*}
\text{all clear} \\
\text{I’m going north}
\end{align*}

$p(\text{ñ} | \text{a})$

\begin{align*}
\text{listo} \\
\text{voy al norte}
\end{align*}
A statistical MT problem

How do we induce a translation model?
A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

[I’m going north]

[Davidson 67. Truth and meaning.]
A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

I’m going north

[Images showing different scenarios with green and grey blocks and a car moving in different directions, indicating correct and incorrect truth conditions.]
A “semantic MT” problem

The meaning of an utterance is given by its truth conditions

→ the distribution over states in which it is uttered

A “semantic MT” problem

The meaning of an utterance is given by its truth conditions
→ the distribution over states in which it is uttered
→ the belief it induces in listeners

I’m going north

[Frank et al. 09. Informative communication in word production and word learning.]
Translating via belief

1.0  2.3
-0.3  0.4
-1.2  1.1
Translating via belief

\[
\begin{pmatrix}
1.0 & 2.3 \\
-0.3 & 0.4 \\
-1.2 & 1.1
\end{pmatrix}
\]

in the intersection
<table>
<thead>
<tr>
<th>1.0</th>
<th>2.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>-1.2</td>
<td>1.1</td>
</tr>
</tbody>
</table>

I’m going north
Translating via belief

\[
\begin{align*}
argmin_a \quad \delta(p(x|\theta) \| p(x|a))
\end{align*}
\]

[A, Dragan and Klein 17. Translating neuralese.]
Computing beliefs

agent policy

agent model

actions & messages
Computing beliefs

human policy

human model

actions & messages
Translating via belief

\[
\begin{pmatrix}
1.0 & 2.3 \\
-0.3 & 0.4 \\
-1.2 & 1.1
\end{pmatrix}
\]

I'm going north
Reference games

1.0  2.3  
-0.3  0.4  
-1.2  1.1
Evaluation: translator-in-the-loop

orange bird with black face
Example translations: image classification

large bird, black wings, black crown

small brown, light brown, dark brown
Experiment: image references

- Chance:
  - Neuralese → English: 50
  - English → Neuralese: 50

- Co-occurrence:
  - Neuralese → English: 57
  - English → Neuralese: 55

- Belief:
  - Neuralese → English: 75
  - English → Neuralese: 60
Example translations: driving

at goal
done
left to top
you first
following
going down

going in intersection
proceed
going
Experiment: driving game

- No communication: 1.93
- Co-occurrence: 1.35
- Belief: 1.49
- Score: 1.54

Learned agent → Learned agent
Beyond communication

≈ orange belly

1.0 2.3
-0.3 0.4
-1.2 1.1

...
Beyond communication

\[ \begin{array}{cc}
1.0 & 2.3 \\
-0.3 & 0.4 \\
-1.2 & 1.1 \\
\end{array} \approx \text{orange belly} \]
Beyond communication

\[
\begin{pmatrix}
1.0 & 2.3 \\
-0.3 & 0.4 \\
-1.2 & 1.1
\end{pmatrix}
\approx \text{orange belly}
\]
Beyond communication

\[
\arg\min_a \delta(p(x|\theta) \parallel p(x|a))
\]

\[
\begin{pmatrix}
1.0 & 2.3 \\
-0.3 & 0.4 \\
-1.2 & 1.1
\end{pmatrix}
\quad \approx \quad \text{orange belly}
\]
1. Represent “meanings” of learned representations and features as distributions over model inputs, and generate explanations by summarizing these distributions with language.
Explanation as (program) induction

Evan Hernandez  Jesse Mu

(thanks to Sarah Schwettman, David Bau, Antonio Torralba)
Explaining representations

\[ \argmin_a \delta(p(x|\theta) \| p(x|a)) \]

\[ p(x|\theta) \]

\[ \approx \text{orange belly} \]
Explaining individual neurons
Explaining individual neurons

[Bau et al. 16. Network Dissection.]
Labeling neurons with categories

[Bau et al. 16. Network Dissection.]
Labeling neurons with categories

argmax\[\text{sim}(\cdot \rightarrow \text{a}, \cdot \rightarrow \text{a})\]

[Bau et al. 16. Network Dissection.]
Labeling neurons with categories

argmax \ \sim(\cdot \rightarrow \text{lamp})

[Bau et al. 16. Network Dissection.]
Neurons as “words” of neuralese

[Bau et al. 16. Network Dissection.]
Are categories enough?

$$\text{argmin}_{a \in \{\text{arena, bank, \ldots}\}} \delta(p(\cdot | b), p(\cdot | a))$$
Are categories enough?

\[
\text{argmin } a \in \{\text{arena, bank, …}\}\\
\delta(p(\cdot | x), p(\cdot | [a])
\]

We see diminishing returns after length 10, so we conduct the rest of our analysis with length 10 logical forms. The increased explanation quality (see Appendix A) suggests that our compositional explanations indeed detect behavior beyond simple atomic labels: for each neuron as we increase the maximum formula length, we learn a perceptual abstraction, or firing for unrelated concepts? Both happen: we manually inspected a random 50 (57%) learn a perceptual abstraction (i.e. the logical form that is also lexically coherent, in that the primitive is a vertical rails detector, but we have no annotations of vertical rails in Broden (Figure 5b). There are many “dead” neurons in the model, and some neurons fire more often than others; we limit our analysis to neurons that activate reliably across the dataset, defined as being active at least 500 times (5%) across the 10K examples probed. For our composition operators, we keep heuristics \[\text{Penn Treebank part of speech tags (labeled by SpaCy https://spacy.io/)}\]; IoU scores steadily increase as max formula length increases. Figure 3 (left) plots the distribution of IoU scores for the best concepts found for prediction. Instead, we analyze the MLP component, probing the 1024 neurons of the penultimate hidden layer for sentence-level explanations, so our inputs, let the hidden layer for sentence-level explanations, so our inputs, let the

4 Do neurons learn compositional concepts?

for each neuron as we increase the maximum formula length

10 (12%) have the form (10)

28 (32%) learn a perceptual abstraction, or firing for unrelated concepts? Both happen: we manually inspected a random 50 (57%) learn a perceptual abstraction, or firing for unrelated concepts? Both happen: we manually inspected a random 50 (57%) learn a perceptual abstraction, or firing for unrelated concepts? Both happen: we manually inspected a random...
We use the SNLI validation set as our probing dataset (10K examples). As our features, we take the unrelated concepts. The 88 “meaningful” neurons fell into 3 categories (examples in Figure 5; more abstractions, or firing for unrelated concepts? Both happen: we manually inspected a random operator. Given a word feature neurons might fire for groups of words with similar meanings, we introduce the unary N operator. Given a word feature RNN features are learned early and are often quite distant from the final sentence representation used for prediction. Instead, we analyze the MLP component, probing the 1024 neurons of the penultimate hidden layer for sentence-level explanations, so our inputs are premise-hypothesis pairs. Additionally, to detect whether models are using lexical overlap for each neuron as we increase the maximum formula length that is also lexically coherent, in that the primitives are specialized, which we call Figure 4 shows an example of a for each neuron as we increase the maximum formula length that is also lexically coherent, in that the primitives are specialized, which we call Figure 4 shows an example of a for each neuron as we increase the maximum formula length that is also lexically coherent, in that the primitives are specialized, which we call Figure 4 shows an example of a for each neuron as we increase the maximum formula length that is also lexically coherent, in that the primitives are specialized, which we call Figure 4 shows an example of a for each neuron as we increase the maximum formula length that is also lexically coherent, in that the primitives are specialized, which we call Figure 4 shows an example of a for each neuron as we increase the maximum formula length that is also lexically coherent, in that the primitives are specialized, which we call Figure 4 shows an example of a for each neuron as we increase the maximum formula length.
We use the SNLI validation set as our probing dataset (10K examples). As our features, we take the
we conduct the rest of our analysis with length 10 logical forms. The increased explanation quality
some meaningful combination of concepts, while
suggests that our compositional explanations indeed detect behavior beyond simple atomic labels:
others; we limit our analysis to neurons that activate reliably across the dataset, defined as being
activations, instead of dynamically thresholding we simply define our neuron masks
as measured by their cosine distance in GloVe embedding space \([\operatorname{cosine distance}]\). Given a word feature
neurons might fire for groups of words with similar meanings, we introduce the unary \(N\)
25%, 50%, or 75% overlap, as measured by IoU between the unique words.

appears in the premise or hypothesis. Additionally, to detect whether models are using lexical overlap
for prediction. Instead, we analyze the MLP component, probing the 1024 neurons of the penultimate
hidden layer for sentence-level explanations, so our inputs
RNN features are learned early and are often quite distant from the final sentence representation used

\(\delta(p(\cdot|l^o)), p(\cdot|l^a)\)

argmin\(a \in \{\text{arena, bank, \ldots}\}\)
From categories to logical forms

argmin \( a \in \{\text{arena, bank, arena or bank, arena and bank, not bank, arena and not...}\} \)

\[ \delta(p(\cdot | a), p(\cdot | a)) \]

[Mu & A 20. Compositional Explanations of Neurons.]
From categories to logical forms

(a) inputs $x$

(b) neuron $f_{483}(x)$

c) neuron masks $M_{483}(x)$

[Mu & A 20. Compositional Explanations of Neurons.]
From categories to logical forms

(a) inputs $x$

(b) neuron $f_{483}(x)$

(c) neuron masks $M_{483}(x)$

(d) concepts $C(x)$

While we explain this neuron’s behavior in human-understandable terms? we generate an explanation via beam search, starting with an inventory of primitive concepts (d), then compare binary neuron masks and concepts with the Intersection over Union score (IoU, or binary masks NetDissect explains the neuron image segmentation masks; for the description activation on a set of concrete inputs (e.g. ResNet-18 [model trained for a language processing task. Now consider an individual neuron be a prefix of a convolutional network trained for image classification or a sentence embedding description.

Consider a neural network model 2 Generating compositional explanations

The quality and interpretability of these learned concepts relate to model performance? Third, can we use incrementally building up more complex logical forms (e). We attempt to maximize the IoU score of Figure 1: Given a set of inputs (a) and scalar neuron activations (b) converted into binary masks (c),

From categories to logical forms
While water (Figure 1d). Given some measure we explain this neuron’s behavior in human-understandable terms? The intuition underlying our approach is shared with the NetDissect procedure of Bau et al. Jaccard similarity; Figure 1f): then compare binary neuron masks and concepts with the Intersection over Union score (IoU, or thresholding above 0), or by dynamically thresholding above a neuron-specific percentile. We can generate an explanation via beam search, starting with an inventory of primitive concepts (d), then incrementally building up more complex logical forms (e). We attempt to maximize the IoU score of binary masks

Formally, assume we have a space of pre-defined atomic description here we describe a generalized version. The core of this intuition is that a good explanation is a quality and interpretability of these learned concepts relate to model performance? Third, can we use abstractions are e.g. gender and lexical overlap) are not only learned, but reified in individual neurons. Compositional explanations allow users to predictably manipulate model behavior: we can generate crude “copy-paste” adversarial examples based on inserting words and image

Consider a neural network model (c) inputs x

water IoU .14 river IoU .08 blue IoU .006 (water OR river) AND NOT blue IoU .15 (water OR river) AND NOT blue IoU .16 (e) logical forms L(x)

c) neuron masks M_{483}(x)

(b) neuron f_{483}(x)
Generate a diagram of figure 1, showing the logical forms and neuron masks along with the inputs.
We use the SNLI validation set as our probing dataset (10K examples). As our features, we take the sample of 128 neurons in the network and their length 10 explanations, and found that suggests that our compositional explanations indeed detect behavior beyond simple atomic labels: for each neuron as we increase the maximum formula length activations, instead of dynamically thresholding we simply define our neuron masks.

Figure 4 shows an example of a sample of 128 neurons in the network and their length 10 explanations, and found that.

For our composition operators, we keep RNN features are learned early and are often quite distant from the final sentence representation used. Additionally, to detect whether models are using lexical overlap others; we limit our analysis to neurons that activate reliably across the dataset, defined as being 25%, 50%, or 75% overlap, as measured by IoU between the unique words.

There are many "dead" neurons in the model, and some neurons fire more often than in the dataset. For each of these we create 2 concepts that indicate whether the word or part-of-speech appears in the premise or hypothesis. Additionally, to detect whether models are using lexical overlap others; we limit our analysis to neurons that activate reliably across the dataset, defined as being 25%, 50%, or 75% overlap, as measured by IoU between the unique words.

Figure 3 shows the distribution of IoU versus max formula length. The line indicates mean IoU.

For our composition operators, we keep RNN features are learned early and are often quite distant from the final sentence representation used. Additionally, to detect whether models are using lexical overlap others; we limit our analysis to neurons that activate reliably across the dataset, defined as being 25%, 50%, or 75% overlap, as measured by IoU between the unique words.

Figure 3 (left) plots the distribution of IoU scores for the best concepts found. There are many "dead" neurons in the model, and some neurons fire more often than.

Figure 4: NetDissect [bullring]
Compositional explanations

Unit 106 bullring OR pitch OR volleyball court
Compositional explanations

Unit 106 bullring OR pitch OR volleyball court AND (NOT football field) AND (NOT railing)
We can now answer our first question from the introduction: are neurons learning meaningful representations of the world? To answer this question, we start by defining our neuron masks. We conduct the rest of our analysis with length 10 logical forms. The increased explanation quality is evident in Figure 3 (left) by the distribution of IoU scores for the best concepts found; IoU scores steadily increase as max formula length increases. There are many "dead" neurons in the model, and some neurons fire more often than would be expected randomly. In limiting our analysis to neurons that activate reliably across the dataset, we have 128 neurons in the network and their length 10 explanations, and found that 88 "meaningful" neurons fell into 3 categories (examples in Figure 5; more details in Appendix A). Some meaningful combinations of concepts, while others are unrelated. For example, the word "bullring" is a vertical rails detector, but we have no annotations of vertical rails in Broden (Figure 5b). Some concepts are not obviously semantically related. For example, 28 (32%) learn a perceptual specialization of the word "bullring" or "cradle" or "autobus" or "fire escape". Some combinations of concepts that are semantically related produce neurons that are activated for groups of words with similar meanings, we introduce the unary N operator. Given a word feature \( x \), let the \( N \) of \( x \) be \( \text{AND NOT} \) \( x \) \( \cdot \) \( \text{EIGHBORS} \) \( x \) \( \cdot \) \( \text{NEIGHBORS} \) \( x \), which we call \( \text{NEIGHBORS} \) \( x \) \( \cdot \) \( \text{EIGHBORS} \) \( x \) \( \cdot \) \( x \). There are many "dead" neurons in the model, and some neurons fire more often than would be expected randomly. In limiting our analysis to neurons that activate reliably across the dataset, we have 128 neurons in the network and their length 10 explanations, and found that 88 "meaningful" neurons fell into 3 categories (examples in Figure 5; more details in Appendix A). Some meaningful combinations of concepts, while others are unrelated. For example, the word "bullring" is a vertical rails detector, but we have no annotations of vertical rails in Broden (Figure 5b). Some concepts are not obviously semantically related. For example, 28 (32%) learn a perceptual specialization of the word "bullring" or "cradle" or "autobus" or "fire escape". Some combinations of concepts that are semantically related produce neurons that are activated for groups of words with similar meanings, we introduce the unary N operator. Given a word feature \( x \), let the \( N \) of \( x \) be \( \text{AND NOT} \) \( x \) \( \cdot \) \( \text{EIGHBORS} \) \( x \) \( \cdot \) \( \text{NEIGHBORS} \) \( x \), which we call \( \text{NEIGHBORS} \) \( x \) \( \cdot \) \( \text{EIGHBORS} \) \( x \) \( \cdot \) \( x \). There are many "dead" neurons in the model, and some neurons fire more often than would be expected randomly. In limiting our analysis to neurons that activate reliably across the dataset, we have 128 neurons in the network and their length 10 explanations, and found that 88 "meaningful" neurons fell into 3 categories (examples in Figure 5; more details in Appendix A). Some meaningful combinations of concepts, while others are unrelated. For example, the word "bullring" is a vertical rails detector, but we have no annotations of vertical rails in Broden (Figure 5b). Some concepts are not obviously semantically related. For example, 28 (32%) learn a perceptual specialization of the word "bullring" or "cradle" or "autobus" or "fire escape". Some combinations of concepts that are semantically related produce neurons that are activated for groups of words with similar meanings, we introduce the unary N operator. Given a word feature \( x \), let the \( N \) of \( x \) be \( \text{AND NOT} \) \( x \) \( \cdot \) \( \text{EIGHBORS} \) \( x \) \( \cdot \) \( \text{NEIGHBORS} \) \( x \), which we call \( \text{NEIGHBORS} \) \( x \) \( \cdot \) \( \text{EIGHBORS} \) \( x \) \( \cdot \) \( x \).
While composition of Broden annotations explains a majority of the abstractions learned, there is still considerable unexplained behavior. The remaining behavior could be due to noisy activations, unrelated disjunctions of concepts, and identify more interesting cases of behavior (e.g. however, allows us to more precisely characterize whether these neurons detect abstractions or specialization (c), and 

Figure 5: Image classification explanations categorized by

<table>
<thead>
<tr>
<th>Unit</th>
<th>Semantic neuron</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>192</td>
<td>skyscraper OR lighthouse OR water tower</td>
<td>0.06</td>
</tr>
<tr>
<td>310</td>
<td>sink OR bathtub OR toilet</td>
<td>0.16</td>
</tr>
</tbody>
</table>
While composition of Broden annotations explains a majority of the abstractions learned, there is still considerable unexplained behavior. The remaining behavior could be due to noisy activations, neuron misclassifications, or detection of concepts absent from Broden.

The observation that IoU scores do not increase substantially past length 10 corroborates the finding of polysemanticity (d). For clarity, logical forms are length 5 explanations. For each neuron, we show the explanation (e.g., Figure 6: NLI length 5 explanations. For each neuron, we show the explanation (e.g., Figure 5: Image classification explanations categorized by...)}
While composition of Broden annotations explains a majority of the abstractions learned, there is neuron misclassifications, or detection of concepts absent from Broden. However, allows us to more precisely characterize whether these neurons detect abstractions or specialization (c), and Figure 6: NLI length 5 explanations. For each neuron, we show the explanation (e.g. and specialization (c), and (Figure 6)

**Unit 483** *(water OR river) AND NOT blue*  
IoU 0.13

**Unit 432** *(attic AND (NOT floor) AND (NOT bed))*  
IoU 0.15
While composition of Broden annotations explains a majority of the abstractions learned, there is still considerable unexplained behavior. The remaining behavior could be due to noisy activations, unrelated disjunctions of concepts, and identify more interesting cases of behavior (e.g., abstraction (lexical and perceptual) and specialization (c), and (d) specialization). For clarity, logical forms are length (a) abstraction (lexical and perceptual) and (b) abstraction (perceptual only) entail neutral sensitive.

<table>
<thead>
<tr>
<th>Unit 314</th>
<th>operating room OR castle OR bathroom</th>
<th>IoU 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unit 439</th>
<th>bakery OR bank vault OR shopfront</th>
<th>IoU 0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Explainability and model accuracy

**Vision**

- Accuracy when firing vs. IoU

**NLI**

- Accuracy when firing vs. IoU
Natural language inference

Hypothesis DOES NOT contain man AND premise contains man

<table>
<thead>
<tr>
<th>IoU</th>
<th>(w_{\text{entail}})</th>
<th>(w_{\text{neutral}})</th>
<th>(w_{\text{contra}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.123</td>
<td>-0.046</td>
<td>-0.021</td>
<td>0.040</td>
</tr>
</tbody>
</table>

**Pre**   A guy pointing at a giant blackberry.
**Hyp**   A woman tearing down a giant display.

Act 29.31 True **contra** Pred **contra**

**Pre**   A man in a hat is working with...flowers.
**Hyp**   Women are working with flowers.

Act 27.64 True **contra** Pred **contra**
Adversarial examples

aqueduct OR viaduct
OR cloister-indoor

bridge OR viaduct
OR aqueduct

viaduct

0.48

washed OR laundromat
OR viaduct

0.46

308

washed OR laundromat
OR viaduct

0.36

347

378

347

26
Adversarial examples

ResNet18
AlexNet
ResNet50
DenseNet161

aqueduct OR viaduct
OR cloister-indoor

26

bridge OR viaduct
OR aqueduct

viaduct

378

washer OR laundromat
OR viaduct

308

ResNet18
forest path
AlexNet
forest path
ResNet50
forest path
DenseNet161
forest path

0.48

0.46

0.36

347

forest path

forest path

forest path

forest path
Adversarial examples

Aqueduct OR viaduct
OR cloister-indoor

Bridge OR viaduct
OR aqueduct

Washer OR laundromat
OR viaduct

ResNet18
AlexNet
ResNet50
DenseNet161

Forest path
Forest path
Forest path
Forest path

Viaduct
Viaduct
Viaduct
Laundromat

0.48
0.46
0.36
1. Represent “meanings” of learned representations and features as distributions over model inputs, and generate explanations by summarizing these distributions with language.

2. Compositionality matters! Lots of interpretable behaviors can’t be explained with a fixed vocabulary of primitives.
Are logical forms enough?

- sides of cylinders
- brown animals and foods
- bird legs and beaks
- high-contrast diagonals
Natural language neuron descriptions

1. This is the side of containers. [A1WS884S10SLO4]
2. The space on the sides of cylinders [A2CFSXJDV50DVF]
3. outer edges of cylindrical shapes [A9V0V2HY30WNG]

1. Human and animal faces and bodies, and vegetables. [ACJ33ZICZI18Q]
2. about animals eye [AH6W2VYAEZNM8]
3. These are animal heads. [A1WS884S10SLO4]
4. They contain pointed features like noses or fruit stems. [A1DP4BJKMEF6LS]

1. Grass and a flower [ACJ33ZICZI18Q]
2. areas of vegetation in a background [A2U1NN8QDH1YVC]
3. These are spots of grass. [A1WS884S10SLO4]
4. about animal [AH6W2VYAEZNM8]
1. Human and horse faces are shown. [A1QAQ72HQ921UG]
2. Grass, the side of a horse head, and the human face beside a horse. [A2XVDB1OXWGTN9]
3. Faces from horses and humans, also leather belts from halters [AX7K5UODLEK72]

1. blue, red and green parts in clothing and objects [AX7K5UODLEK72]
2. This is a saddle. [A1WS884S10SLO4]
3. They are colored basic shapes. [A1UOGYZFJF4BKW]

1. V shaped objects and parts of clothes [AX7K5UODLEK72]
2. These are chins. [A1WS884S10SLO4]
3. Chins texts and wheels are shown. [A1QAQ72HQ921UG]
Language-based control of GANs

increase activation of neurons whose description contains jockey
Language-based control of GANs

increase activation of neurons whose description contains brown
Conclusion
Explain everything!

\[ \text{argmin} \quad a \in \text{English} \quad \delta(p(x|\cdot), p(x|a)) \]
Green and yellow animals, a yellow smiley face, and a firefighter's head and jacket. 
The heads of animals about animals foods in the packet Dog and fox eyes. This is a animal head. Body part of the birds The bottom portions of faces of animals. Dog heads with black and white. Anything that has the color blue in it. These are animal heads. These are animal heads. about animals This is showing both parts of animals and parts of wheels. The color white These are diagonal lines. Shifting contrast colors, either light–dark–light or dark–light–dark. Doors, windows, and see-through objects. Turtle shells and regular overlapping patterns are shown. Red clothing, vehicles, plants, and objects. Human skin The black areas are highlighted in the images. The images show body coverings of animals including fur, feathers, hair, and claws. It shows an image that has a bit of white in it. These are flowers and animal fur. This regions is that is being highlighted are spots. eyes and mouth face of all animals This is fruit and other circles. Face of dogs Dog faces and bodies. The face area is highlighted in the images. eyes and beak People's faces and other body parts. Green grass, plants, and objects. This is black and white grids. yellow spots surrounded by uniform colors body of the dogs arches all the above are in green color Honeycomb, grid, and keyboard button patterns. This is text. The regions depict lines from a center. They are the west or 9 o'clock sections of circular objects that contain concentric rings. This is the area above dots. They are brownish fur. Texts and blue or yellow areas are shown. This is very natural area of cute butterfly. Objects with curved edges. Core hours are set the times when everyone must be outside in the office. The rest of the working hours is flexible. All images are made up of the colors red and white. They are circular objects. Blue colored objects. Animal skins are in the image. The shiny white part of various objects. The shiny white part of various objects and arches. All objects are rounded.
The next challenge: **pragmatic** explanation

Green and yellow animals, a yellow smiley face, and a firefighter's head and jacket. The heads of animals about animals foods in the packet Dog and fox eyes. This is a animal head. Body part of the birds The bottom portions of faces of animals. Dog heads with black and white. Anything that has the color blue in it. These are animal heads. These are animal heads. This is showing both parts of animals and parts of wheels. The color white These are diagonal lines. Shifting contrast colors, either light–dark–light or dark–light–dark. Doors, windows, and see-through objects. Turtle shells and regular overlapping patterns are shown. Red clothing, vehicles, plants, and objects. **Human skin** The black areas are highlighted in the images. The images show body coverings of animals including fur, feathers, hair, and claws. It shows an image that has a bit of white in it. These are flowers and animal fur. This regions is that is being highlighted are spots. eyes and mouth face of all animals This is fruit and other circles. Face of dogs Dog faces and bodies. The face area is highlighted in the images. eyes and beak People's faces and other body parts. Green grass, plants, and objects. This is black and white grids. yellow spots surrounded by uniform colors body of the dogs arches all the above are in green color Honeycomb, grid, and keyboard button patterns. All objects are rounded. This is text. The regions depict lines radiating from a center. They are the west or 9 o'clock sections of circular objects that contain concentric rings. This is the area above dots. They are brownish fur. Texts and blue or yellow areas are shown. This is very natural area of cute butterfly. Objects with curved edges. Core hours are set the times when everyone must be outside in the office. The rest of the working hours is flexible. All images are made up of the colors red and white. They are circular objects. Blue colored objects. Animal skins are in the image. The shiny white part of various objects and animals. They are the midsections of animals or objects. Eyes of various animals and various objects and animals. They are grids and vertical lines.
The next challenge: \textit{pragmatic} explanation

Grice's conversational maxims:

1. Say what is true
2. Say no more than is required
3. Be relevant
4. Avoid ambiguity

[Grice 1975]
The next challenge: *pragmatic* explanation

What can we learn from language users about how to explain models effectively?

Grice’s conversational maxims:

1. Say what is true
2. Say no more than is required
3. Be relevant
4. Avoid ambiguity

[Grice 1975]
1. Represent “meanings” of learned representations and features as distributions over model inputs, and generate explanations by summarizing these distributions with language.

2. Compositionality matters! Lots of interpretable behaviors can’t be explained with a fixed vocabulary of primitives.

3. “Natural” language matters: real networks are probably too complex for logical explanations alone.
Thank you!

Researchers

Evan Hernandez

Jesse Mu

Sponsors

NVIDIA