Towards Automatic Concept-based Explanations

Amirata Ghorbani  James Wexler  James Zou  Been Kim
Existing methods

- A trained model

It's a cab!
Existing methods

- A trained model

Why do you think it’s a cab?  
What is a cab in your eyes?

It’s a cab!
Existing methods

- Local(Instance-wise) methods ⇒ Most important features of the input

Why do you think it’s a cab?
What is a cab in your eyes?

It’s a cab!
Existing methods

- Local(Instance-wise) methods $\Rightarrow$ Most important features of the input

Why do you think it’s a cab?
What is a cab in your eyes?
Existing methods

- Local(Instance-wise) methods $\Rightarrow$ Most important features of the input

Why do you think it’s a cab?
What is a cab in your eyes?

Wheel? Window?
Existing methods

- Global (label-wise) ⇒ Most important features of the class

Why do you think it’s a cab?
What is a cab in your eyes?

- It’s a cab!
- Cabs have wheels
- Cabs are yellow
- Cabs are on asphalt roads
Existing methods

- Global (label-wise) ⇒ Most important features of the class
- TCAV tests queries

It’s a cab!
Cabs have wheels? Cabs are yellow? Cabs are on asphalt roads?

Why do you think it’s a cab?
What is a cab in your eyes?
• In what follows:
  ○ Review Concept Activation Vectors (CAVs)
  ○ Review the TCAV method
  ○ Introduce Concept Discovery in deep neural networks
  ○ Introduce ACE method
  ○ Describe ACE experiments and results
Concept Activation Vectors (CAVs)

- Define a concept to test ⇒ wheel, asphalt texture, etc.
Concept Activation Vectors (CAVs)

- Define a concept to test ⇒ wheel, asphalt texture, etc
- Choose a bottleneck layer
Concept Activation Vectors (CAVs)

- Define a concept to test ⇒ wheel, asphalt texture, etc
- Choose a bottleneck layer

Examples of the concept
Concept Activation Vectors (CAVs)

- Define a concept to test ⇒ wheel, asphalt texture, etc
- Choose a bottleneck layer

Random examples
Concept Activation Vectors (CAVs)

- Define a concept to test ⇒ wheel, asphalt texture, etc
- Choose a bottleneck layer

Linear binary classifier
Concept Activation Vectors (CAVs)

- Define a concept to test ⇒ wheel, asphalt texture, etc
- Choose a bottleneck layer

Linear binary classifier
How good is the classification?
Concept Activation Vectors (CAVs)

- Define a concept to test ⇒ wheel, asphalt texture, etc
- Choose a bottleneck layer

Random examples
Concept Activation Vectors (CAVs)

- Define a concept to test ⇒ wheel, asphalt texture, etc
- Choose a bottleneck layer
Define a concept to test ⇒ wheel, asphalt texture, etc
Choose a bottleneck layer

Concept Activation Vectors (CAVs)
Concept Activation Vectors (CAVs)

- Test example: Is the concept associates with network’s decision
Test example: Is the concept associated with network's decision?
Concept Activation Vectors (CAVs)

- Hypothesis?
Testing Concept Activation Vectors (TCAV)

- Repeat for bunch of test examples: Concept Cav VS Random Cavs ⇒ Statistical Test
Testing Concept Activation Vectors (TCAV)

- **TCAV score** = Ratio of test examples where
Testing Concept Activation Vectors (TCAV)

- TCAV works for human concepts
  - Good for interpretability
  - A few labeled examples (10-30) are shown to be enough
Testing Concept Activation Vectors (TCAV)

- TCAV works for man-defined concepts
  - Good for interpretability
  - Easy to label a few examples
  - Hard to keep tractable

  - Striped? Horizontally Striped?
    - Black-&-white striped?
Testing Concept Activation Vectors (TCAV)

- TCAV works for man-defined concepts
  - Good for interpretability
  - Easy to label a few examples
  - Hard to keep tractable
  - Striped? Horizontally Striped?
    Black-&-white striped?
Testing Concept Activation Vectors (TCAV)

- TCAV works for man-defined concepts
  - Good for interpretability
  - Easy to label a few examples
  - Hard to keep tractable?
    - Striped? Horizontally Striped? Black-&-white striped?
  - Super-human performance
Testing Concept Activation Vectors (TCAV)

- TCAV works for man-defined concepts
  - Good for interpretability
  - Easy to label a few examples
  - Hard to keep tractable?
    - Striped? Horizontally Striped?
      Black-&-white striped?
  - Super-human performance
  - Concepts are not directly related to image pixels
ACE

TCAV

Saliency Maps
ACE

TCAV

Global (General Behavior)

Saliency Maps

Local (Instance-wise behavior)
ACE

TCAV

Global
Concepts
Human-in-the-loop

Saliency Maps
Local
Pixels
Automatic
ACE

TCAV

Global Concepts

Human-in-the-loop

Best of both world

Global

Concepts = Pixels

Automatic

Saliency Maps

Local

Pixels

Automatic
ACE

TCAV

ACE

Police Van

TCAV score = 0.79

TCAV score = 0.78

TCAV score = 0.77

Saliency Maps
ACE

\[ f_i : \mathbb{R}^n \rightarrow \mathbb{R}^m \quad h_{l,c} : \mathbb{R}^m \rightarrow \mathbb{R} \]

Test CAVs
Concept Discovery

- Inputs:
  - A trained model
  - A target class
Concept Discovery

- Inputs:
  - A trained model
  - A target class
  - A bottleneck layer
Concept Discovery

- First step is to discover a class’s concepts e.g. For police van: wheel, sky, asphalt, etc
Concept Discovery

- First step is to discover a class’s concepts ➔ e.g. For police van: wheel, sky, asphalt, etc
- Looking back at CAVs
Concept Discovery

- First step is to discover a class’s concepts e.g. For police van: wheel, sky, asphalt, etc
- Looking back at CAVs highly accurate
Concept Discovery

- First step is to discover a class’s concepts — e.g. For police van: wheel, sky, asphalt, etc
- Looking back at CAVs — highly accurate
Concept Discovery

- First step is to discover a class’s concepts e.g. For police van: wheel, sky, asphalt, etc
- Looking back at CAVs highly accurate
- **Assumption**: Concept examples form clusters in the activation space
Concept Discovery

- First step is to discover a class’s concepts — e.g. For police van: wheel, sky, asphalt, etc
- Looking back at CAVs — highly accurate
- Assumption: Concept examples form clusters in the activation space
- How to find concept examples?
Concept Discovery

- First step is to discover a class’s concepts → e.g. For police van: wheel, sky, asphalt, etc
- Looking back at CAVs → highly accurate
- Assumption: Concept examples form clusters in the activation space
- How to find concept examples?
Concept Discovery

- First step is to discover a class’s concepts e.g. For police van: wheel, sky, asphalt, etc
- Looking back at CAVs highly accurate
- **Assumption**: Concept examples form clusters in the activation space
- How to find concept examples?
  - Can appear several times, once or not at all
  - Appear with different sizes
Concept Discovery

Humans choose concepts → Example concept images → Clustering in activation space

Clustering in activation space → Example concept images → Human interprets concepts

TCAV

ACE
Concept Discovery

- **Idea:** Segment every image with several resolutions ⇒ SLIC
Concept Discovery

- **Idea:** Segment every image with several resolutions ➔ Remove duplicate segments
Concept Discovery

- **Idea:** Segment every image with several resolutions → Remove duplicate segments
- Resize each segment to the network input size
Concept Discovery

- **Idea**: Segment every image with several resolutions
- **Resize each segment to the network input size**

Remove duplicate segments
Concept Discovery

- **Idea**: Segment every image with several resolutions ➔ Remove duplicate segments
- Resize each segment to the network input size ➔ “Resized Patches”
- Map resized patches to activation space
Concept Discovery

- **Idea**: Segment every image with several resolutions → Remove duplicate segments
- Resize each segment to the network input size → “Resized Patches”
- Map resized patches to activation space → Clustering
Concept Discovery

- **Idea:** Segment every image with several resolutions → Remove duplicate segments
- Resize each segment to the network input size → “Resized Patches”
- Map resized patches to activation space → Clustering

Problem: Lots of irrelevant resized patches
Concept Discovery

- **Idea**: Segment every image with several resolutions ➔ Remove duplicate segments
- Resize each segment to the network input size ➔ “Resized Patches”
- Map resized patches to activation space ➔ Clustering with noise removal

**Problem**: Lots of irrelevant resized patches
Concept Discovery

Colors
Concept Discovery

Textures
Concept Discovery

Objects
Concept Discovery

Human related
Concept Discovery

We are running intruder test with human subjects
ACE

\[ f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m \quad h_{l,c} : \mathbb{R}^m \rightarrow \mathbb{R} \]

Test CAVs
ACE

1. Example results: Inception-V3, Mixed-8, Basketball

TCAV score = 0.81

TCAV score = 0.89

TCAV score = 0.88

TCAV score = *

TCAV score = 0.88

TCAV score = 0.50
Example results: Inception-V3, Mixed-8, Drilling Platform

ACE
Example results: Inception-V3, Mixed-8, Volcano

ACE

TCAV score = 0.94

TCAV score = 0.92

TCAV score = 0.89

TCAV score = 0.78

TCAV score = 0.55

TCAV score = 0.88
Experiments

- How to verify ACE?
Experiments

- Concept deletion/addition:
Experiments

● Concept deletion/addition:
  ○ Take a bunch of test images
Experiments

● Concept deletion/addition:
  ○ Take a bunch of test images
  ○ Segment them the same way
Experiments

- Concept deletion/addition:
  - Take a bunch of test images
  - Segment them the same way
  - Assign each patch its NN cluster in activation space
Experiments

- Concept deletion/addition:
  - Take a bunch of test images
  - Segment them the same way
  - Assign each patch its NN cluster in activation space
  - Remove/add patches with concept TCAV score order
Experiments

- Concept deletion/addition:
  - Take a bunch of test images
  - Segment them the same way
  - Assign each patch its NN cluster in activation space
  - Remove/add patches with assigned concept’s TCAV score order

![Addition](image1.png)

![Deletion](image2.png)
Experiments

- Concept deletion/addition:
  - Take a bunch of test images
  - Segment them the same way
  - Assign each patch its NN cluster in activation space
  - Remove/add patches with assigned concept’s TCAV score order
Experiments

- Concept deletion/addition:
  - Take a bunch of test images
  - Segment them the same way
  - Assign each patch its NN cluster in activation space
  - Remove/add patches with assigned concept’s TCAV score order
Experiments

- Concept deletion/addition:
  - Average results for 100 Imagenet classes

![Graphs showing prediction accuracy for SSC and SDC with added or deleted concepts.](image)
Experiments

- Concept stitching experiment:
  - Concepts are discovered as a set of patches
  - We can randomly stitch patches of top-k concepts of each class
Experiments
Thanks!


Code: https://github.com/amiratag/ACE