TorchRay: PyTorch interpretability library for reproducible research

Ruth Fong

University of Oxford (in collaboration with Facebook AI Research)

Open-Source Tutorial for ICCV 2019 XAI Workshop
TorchRay

github.com/facebookresearch/torchray

[ Fong*, Patrick*, & Vedaldi, ICCV 2019 ]
Comparison: TorchRay vs Captum

**TorchRay**

* Supports out-of-the-box methods
* Computer vision (attribution)
* Focus on **reproducible research**: standardized model and benchmarks

**Captum**

* Supports out-of-the-box methods
* Broader support beyond computer vision
* Techniques only
More on motivation

bit.ly/fong19_vgg_interp_tutorial
Follow along in Colab!

bit.ly/torchray_colab_tutorial
Overview

1. How to run attribution methods (colab)
2. How to run benchmark metrics on datasets
3. How to access activations + gradients using Probe objects (colab)
4. Using context managers to implement backprop-based attribution methods (colab)
5. Future work + opportunities to collaborate

Follow along: bit.ly/torchray_colab_tutorial
2. Run benchmark metrics
2. Run benchmark metrics

* By default, expects data to live here:
  - TORCHRAY_DIR/data/datasets/\{imagenet,coco,voc\}
  - Tip: Use symbolic links

```
  ln -s DATASET_DIR TORCHRAY_DIR/data/datasets/
  DATASET_NAME
```

* Run examples/attribution_benchmark.py

* Output stored here: TORCHRAY_DIR/data/
  attribution_benchmarks/ATTRIBUTION_NAME.csv

  gradient,vgg16,voc_2007,0.76281,0.56896
Attribution Methods

1. Gradient
2. Deconvnet
3. Guided backprop
4. Excitation backprop (contrastive + non-contrastive versions)
5. Linear approx
6. RISE
7. Extremal perturbations (ours)

...
Datasets + models

1. VOC + COCO
   * VGG16 and ResNet
   * Ported from original Caffe

2. ImageNet
   * Any model in torchvision

...
Future work + Opportunities to Collaborate

* More models! Self-supervised models, etc.
* More benchmarks! Sanity checks, etc.
* Other techniques! Feature visualization, etc.
* More attribution methods! Your work here!
Thank you!

Email me at ruthfong@robots.ox.ac.uk if you’d like to contribute
TorchRay

github.com/facebookresearch/torchray

PyTorch