A Study on Explanation by Examples for Neural Networks

J.Vikranth Jeyakumar, *Pengrui Quan, Joseph Noor, Yu-Hsi Chen, Mani Srivastava*
vikranth94@g.ucla.edu, mbs@ucla.edu

Other Existing Methods

Perturbation Based

- LIME
  - Creates a local surrogate model
- Anchor
  - If-else rules

Cons
- Lots of hyper-parameters
- Inconsistent over multiple runs

Gradient Attribution Based

- Saliency Maps
- GradCAM ++
- SHAP

Cons
- Mainly designed for images
- Same saliency regions

Deep Neural Networks(DNN) are achieving super-human level performance on a variety of complex tasks involving multiple modalities (Image, text, audio and other sensory data)

But DNNs are Black Boxes

To trust DNNs' decisions, there is human desire for explanations

Motivated externally e.g. by GDPR ("Right to explanation")


Naked Security
GDPR's Right to Explanation: the pros and the cons
Under the General Data Protection Regulation's Right to Explanation a user can ask for an explanation about an algorithmic decision made ...
May 22, 2017

Step 1:

Obtain the activations after last convolutional layer

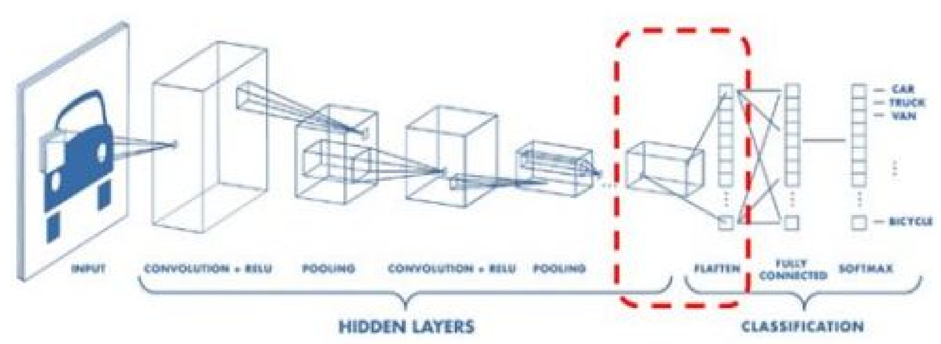
- For the training dataset [$A^k(T)$]
- For the given test sample

Step 2:

Activations are compared using Cosine similarity*

- We choose the top K most similar samples

*Other distance metrics can also be used



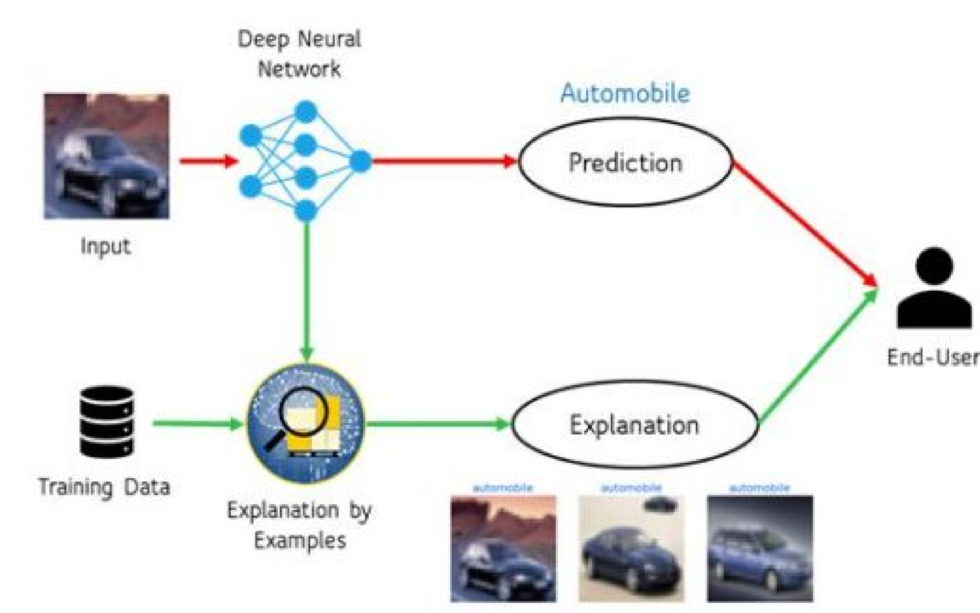
$$examples(x) = \max_{t \in T} \cos(A^k(x), A^k(t))$$

$$= \max_{t \in T} \frac{\sum_{i=1}^{n} A_i^k(x) A_i^k(t)}{\sqrt{\sum_{i=1}^{n} (A_i^k(x))^2} \sqrt{\sum_{i=1}^{n} (A_i^k(t))^2}}$$

Explanation-by-examples requires the Training dataset and the Deep learning model to generate explanation, given a test input

Provides a few key perceptually-relevant items from the training dataset

Explains a DNN decision by looking at which training samples are nearest to a given test input, in the latent space



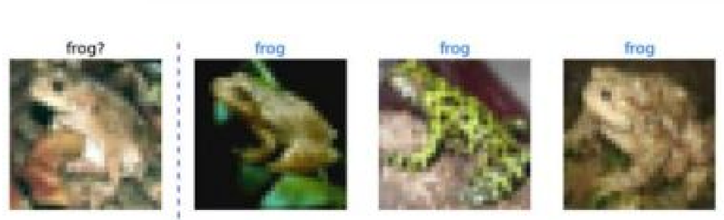
Understandability

We conducted an IRB exempted Amazon Mechanical Turk study comparing different methods to determine the most preferred explanation method for an average end-user

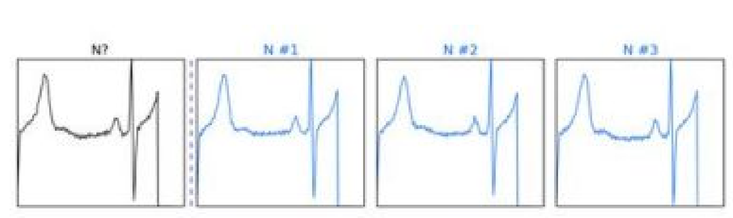
Explanation Method	Image Study	Text Study	Audio Study	ECG Study
LIME	47.7 ± 4.5%	70.4 ± 3.6%	-	-
Anchor	38.9 ± 4.3%	25.8 ± 3.5%	-	-
SHAP	33.7 ± 4.3%	59.9 ± 3.8%	34.7 ± 4.8%	32.8 ± 3.3%
Saliency Maps	39.4 ± 4.3%	-	46.1 ± 5.1%	40.4 ± 3.5%
GradCAM++	50.8 ± 4.5%	-	48.1 ± 5.3%	42.0 ± 3.5%
Explanation by Examples	89.6 ± 2.6%	43.7 ± 3.9%	70.9 ± 4.7%	84.8 ± 2.5%

Results indicate the rate by which users selected a particular method when it is an available explanation, with 95% bootstrap confidence intervals

Explanation by examples can be applied to almost all modalities unlike other existing explanation methods



Explanation: This is a Frog because it looks similar to these images in the training data which are also frogs



Explanation: This is a Normal heartbeat because it looks similar to these samples in the training data which are also normal heartbeats

Cooking with my stepfather

positive?

1 joggin'... with my mom! positive
2 shopping with my bestst! positive
3 glee club party. positive

Explanation: This is a positive sentence because it's similar to the following sentences in the training data which are also positive

Conclusion and Future Work

Explanation by Examples has three key properties:

- Versatility (Works across multiple input modalities)
- Understandability
- Robust to existing adversarial attacks

Given the desirable properties of Explanation-by-Examples method

- We aim to extend our work to explain multi-modal tasks (i.e.) tasks that use multiple modalities simultaneously.
- The job of the adversary is made difficult as it must attack different modalities simultaneously

NESL Networked & Embedded Systems Laboratory

http://www.nesl.ucla.edu/

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake

Existing Adversarial Attacks:

White-box attack:

- Carlini and Wagner Attacks (C&W)

Black-box attack:

- Sign-OPT

Explanation-by-examples detects adversarial examples