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

The Need for Explanation

Other Existing Methods

<table>
<thead>
<tr>
<th>Perturbation Based</th>
<th>Gradient Attribution Based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LIME</strong></td>
<td>• Lots of hyper-parameters</td>
</tr>
<tr>
<td>• Creates a local surrogate model.</td>
<td>• Inconsistent over multiple runs</td>
</tr>
<tr>
<td><strong>Anchor</strong></td>
<td>• SHAP</td>
</tr>
<tr>
<td>• If-else rules</td>
<td>• Mainly designed for images</td>
</tr>
<tr>
<td></td>
<td>• Same saliency regions</td>
</tr>
</tbody>
</table>

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*)

Step 1: Obtain the activations after last convolutional layer
- For the training dataset \( A^T(t) \)
- 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

Step 1:

\[
\text{examples}(x) = \max_{t \in T} \cos(A^T(x), A^T(t)) = \max_{t \in T} \frac{\sum_{i=1}^{K} A^i(x) A^i(t)}{\sqrt{\sum_{i=1}^{K} (A^i(x))^2} \sqrt{\sum_{i=1}^{K} (A^i(t))^2}}
\]

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

<table>
<thead>
<tr>
<th>Explanation Method</th>
<th>Image Study</th>
<th>Text Study</th>
<th>Audio Study</th>
<th>ECG Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIME</td>
<td>47.7 ± 4.5%</td>
<td>70.4 ± 3.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Anchor</td>
<td>38.9 ± 4.3%</td>
<td>25.8 ± 3.5%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SHAP</td>
<td>33.7 ± 4.3%</td>
<td>59.9 ± 3.6%</td>
<td>34.7 ± 4.8%</td>
<td>32.8 ± 3.3%</td>
</tr>
<tr>
<td>Saliency Maps</td>
<td>39.4 ± 4.3%</td>
<td>-</td>
<td>46.1 ± 5.1%</td>
<td>40.4 ± 3.5%</td>
</tr>
<tr>
<td>GradCAM++</td>
<td>50.8 ± 4.5%</td>
<td>-</td>
<td>48.1 ± 5.3%</td>
<td>42.0 ± 3.5%</td>
</tr>
<tr>
<td>Explanation by Examples</td>
<td>89.6 ± 2.6%</td>
<td>43.7 ± 3.0%</td>
<td>70.9 ± 4.7%</td>
<td>84.8 ± 2.5%</td>
</tr>
</tbody>
</table>

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

Conclusion and Future Work
Explaination 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.

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-CW

Explanation-by-examples detects adversarial examples

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