Saliency Oracles for Black Box Predictive Models

Dane Williamson
Department of Computer Science, University of Virginia, Charlottesville, VA 22903
[dw3zn]@virginia.edu

Background

Deep Neural Networks (DNNs) are state-of-the-art learning models which are typically used to make generalisations about potentially complex or otherwise inscrutable data. Despite their widespread usage, they are often criticised for their lack of transparency and their lack of prediction traceability by humans. [4] This lack of transparency and traceability, is a hindrance to the trustworthiness of these models and is likely to impede their implementation in safety-critical domains such as healthcare and autonomous vehicles. In order to foster the essential degree of trust in these models for humans, there must be a higher degree of explainability of the operations of these DNNs. A first step in increasing the explainability of these models is determining the regions of interest (ROI) utilised for a particular classification. These visualized ROI can then provide an intuition behind a DNN classification.

Test Oracles

In computing, software engineering and software testing, a test oracle is a mechanism for determining whether a test has passed or failed. [10] Target label matching is the typical oracle used in determining whether a neural network has made the correct classification on an input. Metamorphic Oracles, i.e. advanced oracles based on adversarial generations on the input provided to the model have been explored in work such as [6] and [9].

Target Labels

The intuitive and commonly used oracle for neural network classifications is to compare the classification output of the model with a predetermined label attached to the input. If the optimal confidence scores output by a classification model correspond with this predetermined label, then the network is considered to have made a correct prediction.

Metamorphic Oracles

A metamorphic test oracle produces new test cases by altering an existing test case, and uses the metamorphic relation between the inputs and the outputs of the system to predict the expected outputs of the produced test cases.[9] A metamorphic relation is defined formally as a conditional rule defining the degree of change expected in the output, when a transformation is applied to the test input. [7]

While these oracles set the precedent of defining an expected output based on the input to the model, none of these oracles requires a model to provide an intuition or explanation for a classification.

Aim

The aim of this study is to develop a framework to improve the explainability of DNN classifications. For this we propose a modification to the standard DNN test oracle. In addition to comparing DNN classification with target labels, we propose comparing salient regions as determined by the network with human-annotated ROI. The saliency oracle will help determine if the explanation provided by the network to support its classification is acceptable.

Method

In order to determine the salient regions for a model classification, we generate overlayed saliency maps on the Fashion-MNIST images. The framework we used to generate these saliency maps is borrowed from [5]. In [5], the researchers develop a framework called Randomized Input-Sampling for Explanation of Black-Box Models (RISE). By computing the measure of overlap between the human-annotated ROI and the salient regions as determined by the model, we seek to provide a more robust test oracle for DNN classifications.

Saliency Oracles

The method of computing a similarity index between the regions of the image emphasized by the network, with those annotated as relevant by the human is here defined as a ***Saliency Oracle***.

Experiments and Results

Implementation

The model developed was implemented in pytorch. The network consisted of one convolutional layer, one max pooling layer and two fully connected layers. The model had training accuracy of $\tilde{9}2$ % and validation accuracy of $\tilde{8}9\%$. This is near enough to benchmark Fashion-MNIST accuracy ($\tilde{9}1\%$) for evaluation with saliency maps to be feasible.

Saliency Maps With RISE

Using the technique outlined in [5] we generate 4000 8x8 binary masks with the probability of each pixel being set to one or zero being 0.5. The binary masks are then up-sampled, cropped and applied to the Fashion-MNIST images, before being provided as inputs to the model. This process determines the saliency, $S_{I,f}$ of each pixel (where I represents the input image and f represents the vector of confidence scores as computed by the black box model)$\lambda \in \Lambda$ through a Monte-Carlo estimation of expected score over all possible masks, M, conditioned on the event that pixel λ is observed:

$$S_{I,f}(\lambda) = \mathbb{E}_M[f(I \bigodot M)|M(\lambda) = 1]. \tag{1}$$

[5]. A sample output of the saliency maps overlayed on a Fashion-MNIST image can be seen in Figure 1. As shown in Figure 1, the saliency maps do not represent what would typically be considered important in classifying an article of clothing. In fact the network seems to emphasize ostensibly irrelevant regions of the image.

Computing Regional Overlap

Intersection over Union (IoU) is also known as Jaccard Index (J.I.) and is the most commonly used metric for comparing the similarity between two arbitrary shapes. The acceptable J.I. value chosen was 0.5. [8] As shown in Figure 2, we annotate the images with regions we deem relevant and then compare the degree of overlap between these annotated regions and the ROIs as dictated by the network. The example shown in Figure 2 is likely to fail the oracle, as there is insufficient overlap to support the classification.

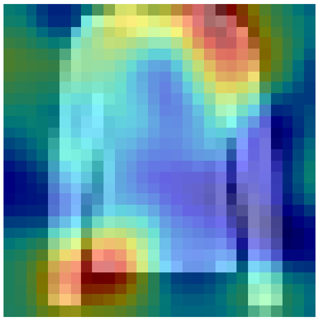

Figure 1. Saliency map emphasizing regions of importance in classifying a ’pullover’.

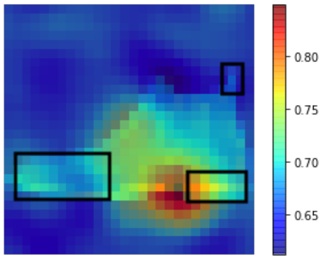

Figure 2. Human annotated ’relevant’ areas in classifying an ankle boot are demarcated with a rectangular box.

Discussion

Given the results shared, we anticipate that these stricter saliency oracles will reduce the benchmark classification accuracy for conventional networks on popular datasets. Possible explanations for the seemingly random nature of the saliency maps generated include :

1. The model was trained to identify from only 10 classes. This may provide an intuition for why the model did not ostensibly discriminate effectively.

2. The images fed to the model were not subjected to sufficient data augmentation in training to enable the model to focus on relevant features.

Future Work

Dataset Annotations

In adding annotations of relevant regions of an image we can borrow from annotation techniques used in popular object detection datasets such as KITTI [1] and MS-COCO [2]. MS-COCO contains 328k images with over 2.5 million labelled instances, while KITTI contains over 12919 images. To annotate a popular image recognition dataset would thus require a significant investment of human effort, which is far beyond the scope of this project. The precedent established by these popular object detection datasets indicates that annotating a sufficiently large dataset is nonetheless achievable despite the obvious expense.

Training Process

The explanations provided by frameworks such as [5] may give an intuition for what steps can be taken to improve model accuracy. These steps may include applying angular rotations to images in training, perhaps adding convolutional layers or training our model on a dataset with a sufficiently large number of classes.

Additionally, we could train our model to focus on the human annotated ROI. In [3], researchers augment their loss function to consist of two parts:

1. The localization loss for bounding box offset prediction
2. The classification loss for conditional class probabilities.

Both parts are computed as the sum of squared errors. We could further augment the loss function used in [3] to train our model so that the saliency maps generated sufficiently intersect with human-annotated ROI.

Code

All the code used in this project can be found in the GitHub repository.

References

[1] Andreas Geiger et al. “Vision meets Robotics: The KITTI Dataset”. In: *International Journal of Robotics Research* 32.11 (Sept. 2013), pp. 1231–1237.

[2] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[3] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *CoRR* abs/1506.02640 (2015). arXiv: 1506.02640. URL: http://arxiv.org/abs/1506.02640.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

[5] Vitali Petsiuk, Abir Das, and Kate Saenko. “RISE: Randomized Input Sampling for Explanation of Black-box Models”. In: *CoRR* abs/1806.07421 (2018). arXiv: 1806.07421. URL: http://arxiv.org/abs/1806.07421.

[6] Anurag Dwarakanath et al. “Metamorphic Testing of a Deep Learning based Forecaster”. In: *CoRR* abs/1907.06632 (2019). arXiv: 1907.06632. URL: http://arxiv.org/abs/1907.06632.

[7] Rohan Reddy Mekala et al. “Metamorphic Detection of Adversarial Examples in Deep Learning Models With Affine Transformations”. In: *CoRR* abs/1907.04774 (2019). arXiv: 1907.04774. URL: http://arxiv.org/abs/1907.04774.

[8] Seyed Hamid Rezatofighi et al. “Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression”. In: *CoRR* abs/1902.09630 (2019). arXiv: 1902.09630. URL: http://arxiv.org/abs/1902.09630.

[9] Adrian Wildandyawan and Yasuharu Nishi. *Object-based Metamorphic Testing through Image Structuring*. 2020. arXiv: 2002.07046 [cs.LG].

[10] Wikipedia. *Test oracle — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Test%20oracle&oldid=1004406321. [Online; accessed 04-March-2021]. 2021.