May 7th, 11:00 AM - 1:00 PM

Explanation Methods for Neural Networks

Jack H. Chen
Portland State University

Christof Teuscher
Portland State University

Explanation Methods for Neural Networks

Jack Chen, Walt Woods
Advisor: Christof Teuscher

Portland State UNIVERSITY

Maseeh College of Engineering and Computer Science
PORTLAND STATE UNIVERSITY

teuscher::Lab
teuscher-lab.com
Emerging Computing Models and Technologies

Abstract

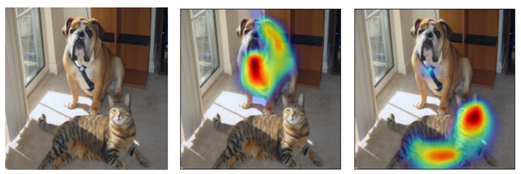
Neural Networks (NNs) have become a basis of almost all state-of-the-art machine learning algorithms and classifiers. While NNs have been shown to generalize well to real-world examples, researchers have struggled to show why they work on an intuitive level. We designed several methods to explain the decisions of state-of-the-art NN classifiers which include using leading explanation methods LIME and Grad-CAM. We analyze regions deemed salient highlighted by these methods, show how these explanations may be misleading, and discuss future directions including methods which construct full color images using adversarial examples rather than heat maps to provide more complete explanations of NN classifiers.

Methods

Explanations methods examined are LIME, Grad-CAM, and adversarial maximization. We explain NN architectures with the Japanese Society of Radiological Technology (JSRT) chest lung nodule CT imaging database and the CIFAR-10 imaging dataset. Additionally, we propose maximizing adversarial attacks as another way of explaining NNs.

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 1: Explanations of target classes via LIME from Ribeiro et al.[1] Images are segmented into superpixels salient to a class.



(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Figure 2: Explanation heatmaps explaining saliency of a different class using Grad-CAM[2].

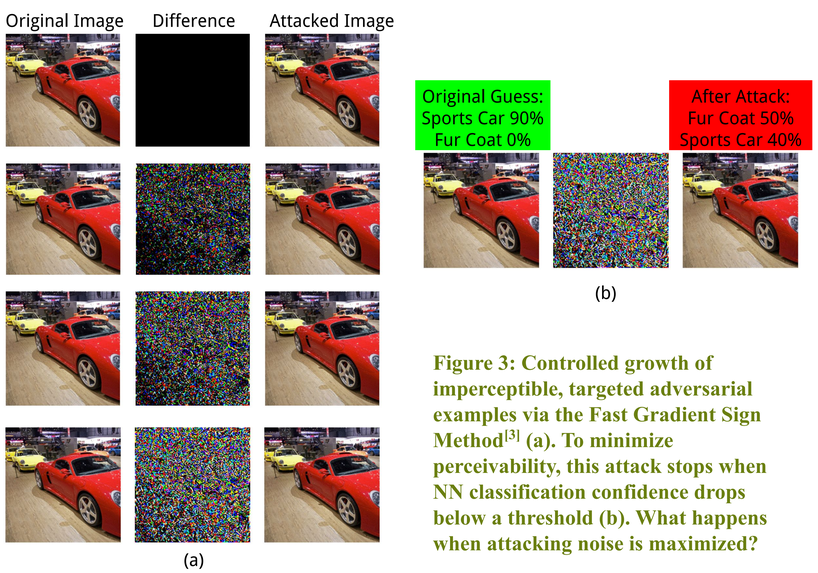


Figure 3: Controlled growth of imperceptible, targeted adversarial examples via the Fast Gradient Sign Method[3] (a). To minimize perceivability, this attack stops when NN classification confidence drops below a threshold (b). What can be inferred about the NN from this method?

JSRT Explanations

We trained a ResNet[4] classifier to identify lung nodules in the JSRT database. We then ran LIME segmentation and Grad-CAM heatmap algorithms to explain areas decided to be lung nodules.

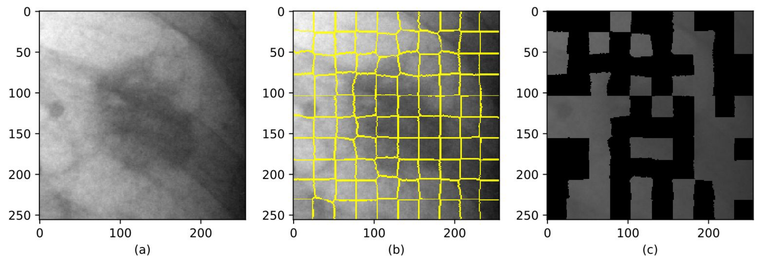


Figure 4: LIME explaining the most relevant segments for a lung nodule via JSRT. What can be inferred about the NN from this method?

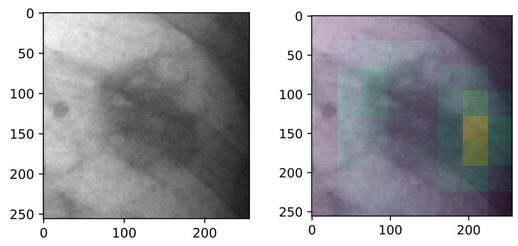


Figure 5: Area of interest for a lung nodule marked with heatmap via Grad-CAM. Most salient region marked in yellow.

CIFAR-10 Explanations

An All Convolutional NN[5] and a ResNet NN were trained for classifying CIFAR-10 images. Explanations are generated via LIME and Grad-CAM before and after applying imperceivable adversarial attacks.

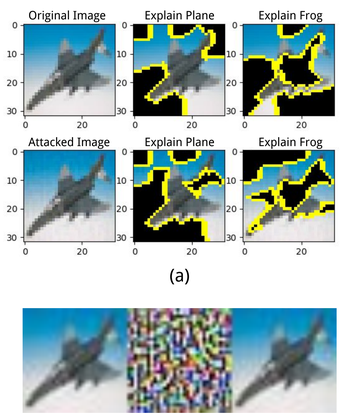


Figure 6: LIME on CIFAR-10 before/after adversarial attack (a). Normalized noise of attack in (b).

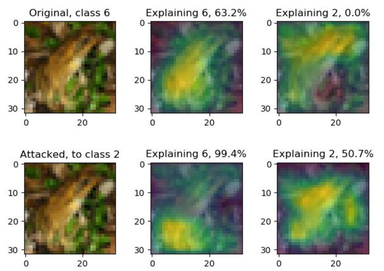


Figure 7: Grad-CAM on CIFAR-10 before/after adversarial attack. Notice shift in heatmap.

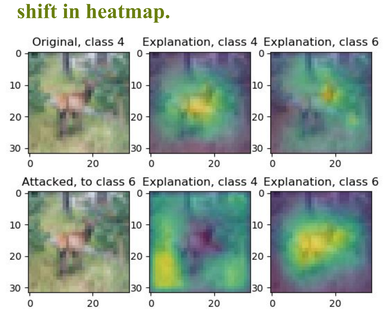


Figure 8: More Grad-CAM explanations

Explanations via Adversarial Attacks

Figure 3(a) shows the evolution of targeted adversarial attacks and stops at threshold. If these attacks have a much larger stopping point, we should see more salient features in an attacked image.

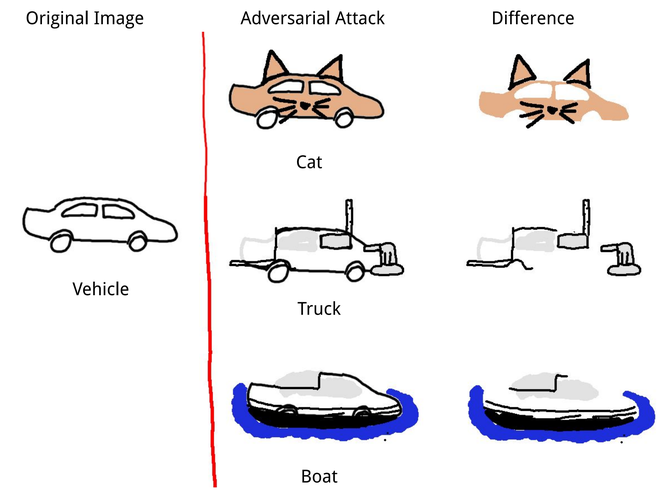


Figure 9: Representations of maximized adversarial attacks providing explanations of important features of classes other than the original class.

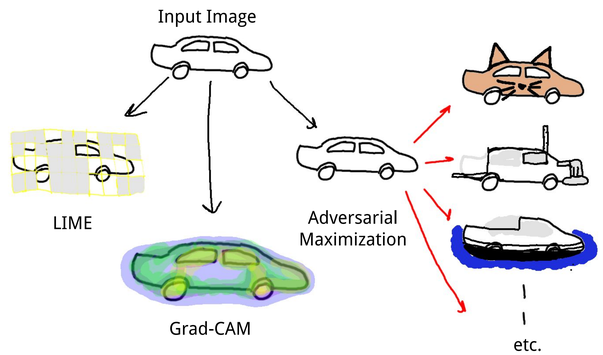


Figure 10: Comparison methods discussed. Adversarial maximization explains other classes rather than directly explaining the input which may provide more insight in NN decision than with only heatmaps.

Conclusion/Results

- LIME shows the most important segments of an input for target class.
- Grad-CAM shows a heatmap of most relevant areas for target class.
- Heatmaps can change with adversarial attacks.
- Adversarial maximization (our method) shows the salient features that changes the decision of the NN.
- Paper pending

References

[1] Ribeiro M., Singh S., Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. 2016.
[2] Li K., Wu Z., Peng K., Ernst J., Fu Y. Tell Me Where to Look: Guided Attention Inference Network. 2017.
[3] Goodfellow I., Schlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. 2015.
[4] He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. 2015.
[5] Springenberg J., Dosovitskiy A., Brox T., Riedmiller M. Striving for Simplicity: The All Convolutional Net. 2015