CSC594 Assignment 4

Clayton Cohn

November 7, 2020

1 Assignment

This assignment entailed two different types of explainable AI (XAI) implementations: Local Interpretable Model-agnostic Explanations (LIME) [3] and Integrated Gradients [2]. The goal was to compare and contrast the two methods across a variety of images and models and then report on the findings.

2 Images

I chose four images to compare and contrast: two images from ImageNet's list of synsets, and two images that were ambiguous and potentially difficult to classify. The two images belonging to the ImageNet synsets were from the "horned_viper" and "military_uniform" synsets. They were chosen as images that should be readily identifiable across all models. Additionally, two images were chosen with the intention of creating a potential challenge for the models: one image was of the professional golfer Phil Mickelson, and the other image was of a neuron and its dendrites.

3 Models

I left the models component of the two notebooks as-is. The models used for both assignments were InceptionV3, VGG16, and ResNet50. Overall, ResNet50 was the best-performing model, as it provided the most sensible classifications for the more difficult images. ResNet classified Phil Mickelson as a "ballplayer" or "golf_ball," while another model classified him as a "parachute." Similarly, ResNet classified the neuron image as a "spider_web," while another model classified it as a "jellyfish." Thus, I felt that ResNet50 was the model that presented the best classifications for the tougher images.

4 Findings

All of the models had no problem identifying the images whose synsets were explicitly present in ImageNet; however, classifying the other images was not as straightforward of a task. It was interesting to see that both approaches (LIME and Integrated Gradients) took very different approaches, even when arriving at the same definitive answer. In the following figures, Figure 1 (LIME) shows that all the model needed to accurately predict the "military_uniform" synset was the Marine's face, neck, and hat. Conversely, in Figure 2 (Integrated Gradients), the model was more interested in the other pieces of military uniform that were present in the photo.

Anecdotally, I did not notice much of a difference in performance: both methods seemed to do the job relatively well. Still, I was curious about what the pros and cons of each method were, so I turned to the original papers ([2] and [3]). Additionally, I also referenced a paper that compared various methods, two of which were LIME and Integrated Gradients [1]. [1] found that LIME consistently underperformed the other approaches (there were four approaches in total, two of which were covered in this assignment). They found that LIME consistently selected the smallest number of important pixels and often failed to recognize important pixels identified by the other models. They actually did not have a single positive thing to say about LIME, which was a bit surprising considering I felt that it did the job fairly well.

Figure 1: *LIME*

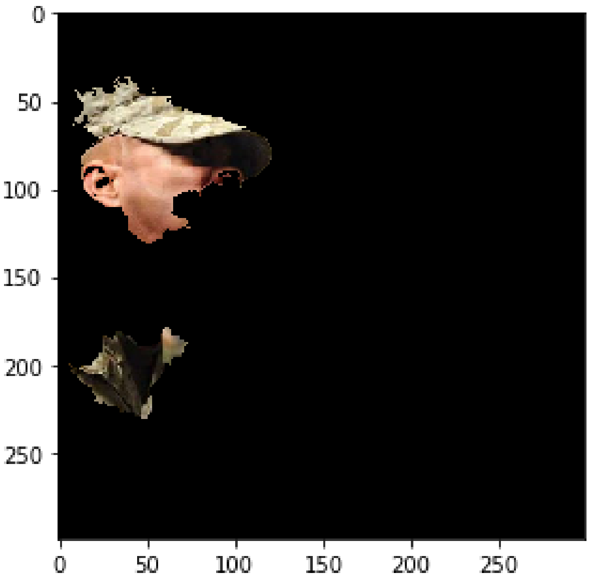
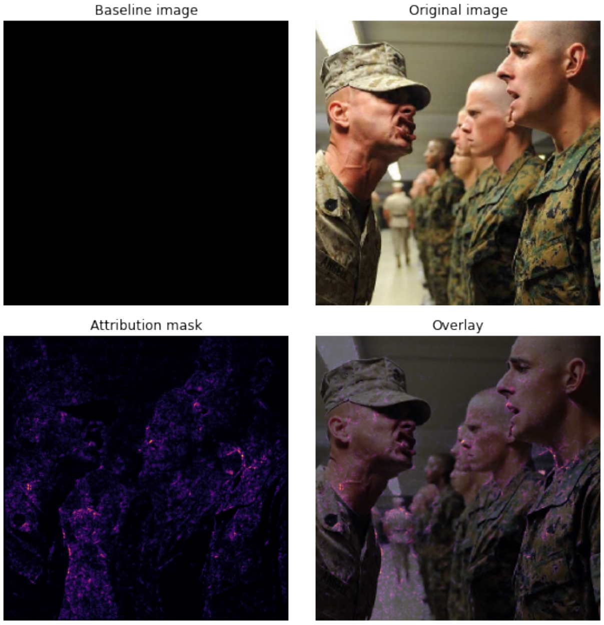


Figure 2: *Integrated Gradients*



For Integrated Gradients, [1] found that this method was somewhere in the middle of the pack. While it did not suffer considerably like some other models when considering inverse masking, its performance rarely beat that of Grad-CAM, which was this study's best-performing model. One thing I did like about this study was how it defined what a "good explainer" was. An explainer, according to [1], should be evaluated on: correctness, consistency, and confidence.

5 Discussion

I think that the concept of XAI is extremely important. Black-box architectures are already causing problems as people are questioning the motives behind many of these models' results—especially if the result is an unpopular one. XAI provides researchers (and the public) the opportunity to peek under the hood and discern why a model comes to a particular conclusion. As far as XAI's viability (i.e., trustworthiness) is concerned, I think that the technology is still in its nascent stages. The fact that both of these models can spit out the same (correct) classification, despite disagreeing on the important pixels, tells me that the technology is still largely open to interpretation. It will be interesting to see what people come up with in the coming years.

Still, I do think that there is a huge takeaway here, which is that we're on the right track. I think that the fact that XAI is being addressed is a testament to our desire to understand these previously inexplicable models. I think it's only a matter of time before the technology is both trusted and ubiquitous. Going forward, I would like to see something similar in the NLP domain. Also, I am curious if there is a way to use this for multilabel classification. For both of my pictures that were not easy to classify, both had multiple items present in the pictures. I would therefore like to see how XAI performs on a multilabel classification task.

References

[1] Gokula Krishnan Santhanam et al. "On Evaluating Explainability Algorithms". In: 2019.

[2] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *arXiv e-prints*, arXiv:1703.01365 (Mar. 2017), arXiv:1703.01365. arXiv: 1703.01365 [cs.LG].

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *arXiv e-prints*, arXiv:1602.04938 (Feb. 2016), arXiv:1602.04938. arXiv: 1602.04938 [cs.LG].