Classifier-Agnostic Crossing of Decision Boundary under the Image Manifold

Anonymous Authors[1]

Abstract

1. Introduction

Understanding the behaviour of complex machine learning predictors is important for debugging and for improving their performance and robustness. Moreover, such understanding can help flag hidden bises in the data (Cramer 2018) and evaluate the fairness of the model (Doshi-Velez and Kim, 2017), and build trust in the system [XX]. However, even if a classifier has 100% accuracy for a given class, it could still fail to understand that particular concept. For example, if all images of a polar bear exhibit a white background, the classifier can learn to make the decision based on the background pixels.

Saliency methods (Simonyan et al., 2013; Springenberg et al., 2014; Selvaraju et al., 2016; Smilkov et al., 2017) are popular tools for interpreting predictors by highlighting the most important features in an image that results in the given classification. While saliency methods have proven useful for a variety of applications [XX], some of them were shown to be agnostic to the predictor model and hence cannot help with debugging (Adebayo et al., 2018). Other methods have relied on local occlusions or perturbations to the input ((**?**), (**?**)) Other methods have focused on counterfactual explanations (Wachter et al., 2017; Goyal et al., 2019), tring to show in what way does the input need to change in order to be assigned a different class.

The work we present here is most related to the work of (Singla et al., 2019; **?**), which from a given starting image, try to find a series of "natural" images that cross the boundary of the black box classifier. One drawback of (Singla et al., 2019) is that they require re-training the generator every time the black-box classifier is changed, which is something that can happen often, especially when debugging. For example, one can imagine using such a visualisation tool to find useful data augmentation strategies. If generator re-training takes long, this can impede its usefulness. Furthermore, this also prohibits the use of state-of-the-art image generators such as (**??**), that require large amounts of computational resources.

In our work we re-define the problem and simplify the formulation of (Singla et al., 2019). More precisely, we assume a fixed, pre-trained generator G that can generate realistic images. Our aim is to further few

2. Method

We assume a black box classifier $f : X \rightarrow R$, an initial input image $x \in X$, and a target class C. Without loss of generality, we assume the classifier is binary and that $f(x) \in [0, 1]$ is the probability of assigning x to target class C. Moreover, compared to (Singla et al., 2019), we also assume a fixed, pre-trained generator $G : Z \rightarrow X$, generating realistic images. We aim to find a trajectory of synthetic images $\mathbf{x} = (x_1, ..., x_T)$ corresponding to latent points $= (z_1, ..., z_T)$, $z_i \in Z$ such that:

- **Monotonicity**: the trajectory monotonically converges towards the target class C: $f(z_t + 1) > f(z_t)$
- **Identity preservation**: latent points z_t are as close as possible to the embedding of the source image $G^{-1}(x)$. For example in computer vision or medical imaging, this helps preserve the identity of the person studied, ensuring only minimal changes are made in order to change the label of the point to target class C.

3. Option 1: Gradient descent along the classifier softmax

$z_{t+1} = z_t + \eta \nabla f(G(z_t))$

4. Option 2: Reguarised optimisation

We define a "step" loss on classifier f, ensuring that

$$L_f(\mathbf{z}) = \sum_{t=1}^{T-1} |f(G(z_{t+1})) - \frac{t}{T}| \qquad (1)$$

$$L_i d(\mathbf{z}) = \sum_{t=1}^{T-1} |G^{-1}(x) - G(z_t)| \qquad (2)$$

$$L(\mathbf{z}) = L_f(\mathbf{z}) + \lambda L_i d(\mathbf{z}) \qquad (3)$$

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451*, 2019.

Joshi, S., Koyejo, O., Kim, B., and Ghosh, J. xgems: Generating examplars to explain black-box models. *arXiv preprint arXiv:1806.08867*, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. Explanation by progressive exaggeration. *arXiv preprint arXiv:1911.00483*, 2019.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.