HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)

doi:10.3233/FAIA220209

Exemplars and Counterexemplars Explanations for Skin Lesion Classifiers

Carlo Metta [a,1], Riccardo Guidotti [b], Yuan Yin [c], Patrick Gallinari [c], and Salvatore Rinzivillo [a]

[a] *ISTI-CNR, Pisa, Italy*
[b] *University of Pisa, Italy*
[c] *Sorbonne Universite, Paris, France*

Abstract. Explainable AI consists in developing models allowing interaction between decision systems and humans by making the decisions understandable. We propose a case study for skin lesion diagnosis showing how it is possible to provide explanations of the decisions of deep neural network trained to label skin lesions.

Keywords. Image classification, Explainable AI, Machine Learning, Skin Lesion Image Classification, Adversarial Autoencoders

1. Introduction

AI based decision support systems have a huge impact in different domains, providing high accuracy predictions and recommendations. Their adoption in high-stake scenarios has raised concerns about the fairness, bias, transparency and dependable decisions taken on the basis of AI suggestions [1]. These concerns are relevant in domains like healthcare: image classification can be applied for purposes ranging from heart disease diagnosis to skin cancer detection [2–4]. Machine learning models are typically *black boxes* hiding the rationale of their behavior, for this reason, research on *black box explanation* has received much attention [5–8]. Our solution [9] is based on the *local model-agnostic* method ABELE. The approach is proved on a classification model for the *ISIC 2019 Challenge* and consists of two modules: *(i)* a CNN model to classify skin lesion images; *(ii)* an explainer that exploits Adversarial AutoEncoder (AAE) to produce images for the explanation. Since it is crucial to have a wide catalog of instances, we developed a progressive growing AAE to maximize the diversification of the generated images. Accurately designing the AAE is crucial for obtaining an explanation based on realistic images.

2. ABELE explainer

The Adversarial Black box Explainer generating Latent Exemplars (ABELE) is a local model agnostic explainer for image classifiers [11]. The explanation is composed of *(i)* a set of *exemplars* and *counter-exemplars*, instances classified with same or different out-

[1] Corresponding Author; E-mail: carlo.metta@isti.cnr.it

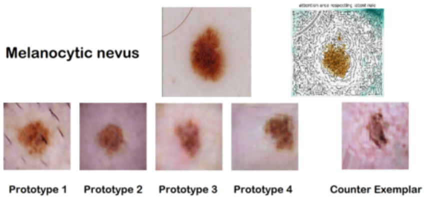

Figure 1. ABELE graphic explanation of a Melanocytic nevus.

come of a given image, *(ii)* a *saliency map* that highlights areas that contribute to the classification and areas that push it toward another class. First, ABELE generates a neighborhood exploiting an AAE [12], then it learns a decision tree on the latent neighborhood providing local decision and counterfactual rules [14], and finally selects and decodes exemplars and counter-exemplars, extracting a saliency map.

3. Case Study

ISIC 2019 is a challenge proposed by the International Skin Imaging Collaboration. The goal is to classify dermoscopic images among different categories. For the classification, we used a classical ResNet, pretrained on Imagenet and fine tuned on ISIC dataset.

We implemented a collection of techniques for successfully training an AAE. Progressive Growing GANs [19] have been introduced to achieve a stable training of generative models. We propose a Progressive Growing AAE (PGAAE): starting with a single block of layers for the generating network we reconstruct low resolution images, then we increase the number of blocks until the network manages images of the desired size.

Denoising autoencoders [20] are a stochastic version which randomly corrupt input image and are proved to learn more robust representations. We augment our PGAAE with noise injection applied to both generator and discriminator. Mini Batch Discrimination [16] is a technique that mitigates collapse of the generator network. Such technique along with the progressive growing structure, helped to avoid the mode collapse .

4. Explanations

The outcome is a compact interface: 1) the original image and the predicted label; 2) a map that emphasizes areas that had a positive or negative contribution to the classification; 3) a set of synthetic prototypes that are classified with the same or different class of the input. Fig. 1 shows a sample explanation. From the map the user can evaluate which parts of the image were relevant for the CNN; the prototypes generated by the AAE enforce the confidence with the black-box decision while the counterexemplar probes the black-box result, by generating an image similar to the input but classified differently.

5. Conclusion

This work is the core of a wider system where the interaction should be further developed by enabling an exploration of the latent space, allowing the user to ask for additional explanations. We design a survey with experts to test different explanations features [10].

References

[1] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, "Meaningful explanations of black box AI decision systems," in *AAAI*. AAAI Press, 2019, pp. 9780–9784.

[2] I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, and J. Ma, "Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images," *Comput. Medical Imaging Graph.*, vol. 88, p. 101843, 2021.

[3] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *ICARCV*. IEEE, 2014, pp. 844–848.

[4] M. Antonie, O. R. Zaïane, and A. Coman, "Application of data mining techniques for medical image classification," in *MDM/KDD*. University of Alberta, 2001, pp. 94–101.

[5] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 93:1–93:42, 2019.

[8] C. Panigutti, A. Perotti, and D. Pedreschi, "Doctor XAI: an ontology-based approach to black-box sequential data classification explanations," in *FAT**. ACM, 2020, pp. 629–639.

[9] C. Metta, R. Guidotti, Y. Yin, P. Gallinari, and S. Rinzivillo, "Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling," in *2021 IEEE Symposium on Computers and Communications (ISCC)*, 2021.

[10] C. Metta, A. Beretta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, and F. Giannotti "Explainable Deep Image Classifiers for Skin Lesion Diagnosis," *arXiv:2111.11863*, 2022.

[11] R. Guidotti, A. Monreale, S. Matwin, and D. Pedreschi, "Black box explanation by learning image exemplars in the latent feature space," in *ECML/PKDD (1)*, ser. Lecture Notes in Computer Science, vol. 11906. Springer, 2019, pp. 189–205.

[12] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow, "Adversarial autoencoders," *CoRR*, vol. abs/1511.05644, 2015.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[14] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intell. Syst.*, vol. 34, no. 6, pp. 14–23, 2019.

[15] H. Thanh-Tung and T. Tran, "Catastrophic forgetting and mode collapse in gans," in *IJCNN*, 2020.

[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. C. and, "Improved techniques for training gans," in *NIPS*, 2016.

[17] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," in *ICLR*, 2017.

[18] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *ICLR*, 2018.

[20] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008.