ICAI 2023
12th International Conference on Applied Informatics
Eger, Hungary, 2023

A comparative study of interpretable image classification models

Adél Bajcsi[a], Anna Bajcsi[a], Zalán Bodó[a], Lehel Csató[a], Szabolcs Pável[ab], Ábel Portik[a], Csanád Sándor[ab], Annamária Szenkovits[a], Orsolya Vas[a]

[a]Faculty of Mathematics and Computer Science, Babeş–Bolyai University, Cluj, Romania

[b]Robert Bosch SRL, Cluj, Romania

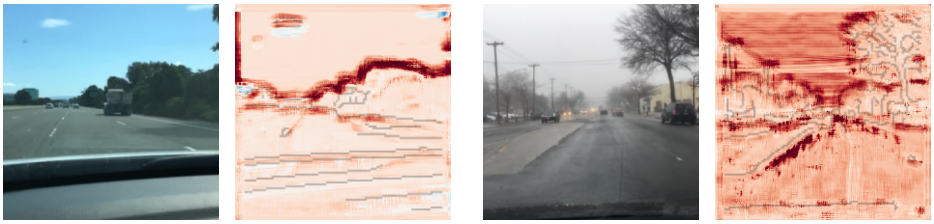
Abstract

Recent advances in deep learning brought a plethora of learning models and network architectures that perform better on a huge variety of problem types and a large variety of data. A distinctive new research direction is to specialize the architecture such that the decisions of the system are explainable; specifically, there is a possibility to identify those parts of the image responsible for the decision. Systems capable to provide explanations for the decisions they made have been called *self-explainable* or *interpretable* models [3, 6].

Self-explaining models are thought to be useful in safety-critical environments, like driver assistance systems and autonomous driving, since they might help with:

- Root cause analysis of failure cases of deep models. Explainable AI can speed up the analysis process by highlighting issues regarding network components/quality issues of the datasets [6].
- Fulfilling regulatory requirements towards Driver Assistance (DA)/Autonomous Driving (AD) systems. Owing to its safety-critical nature, these requirements are often very strict, and the future requirements will be even stricter, setting up the "right to explain" [4].

Apart from safety-critical scenarios, explainable decisions always imply a higher level of trust, therefore, their application can often be more beneficial than that of their non-interpretable versions. For example, when predicting the veracity of a



(a) Clear weather (b) Foggy weather

Figure 1. Sample predictions (BDD100K dataset) obtained using the BagNet-17 model for clear and foggy weather. The darker the region, the more important it was in the process of assigning the given weather class label to the image.

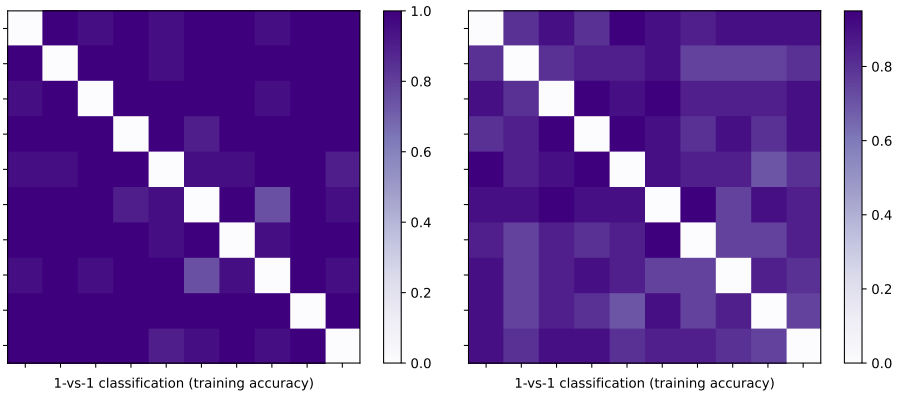
news article in automatic fake news detection systems, justifying the decision made causes a better overall acceptance of these predictions [7].

The recent explainable models can generally be described as two-tier neural network architectures, composed of a feature detection unit – often called *backbone* – followed by a classification or labeling part, with classification being based on the features computed by the backbone. In this paper we study three such architectures: PrototypeDL [5], ProtoPNet [2], and BagNet [1]. PrototypeDL uses an autoencoder to obtain the above-mentioned feature representation, while its second part builds the prototypes to which the images will be compared. The predicted label of an image is calculated based on its distances from the prototype vectors via a fully connected layer. ProtoPNet improves on this model by considering prototypes corresponding to smaller image patches, and constructing a heatmap based on the prototype distances. The third model studied here differs from the previous two in being a fully convolutional neural network (FCN) without explicitly generating prototype vectors. In this case, the FCN can assign importance scores to small image patches, resulting in a heatmap similar to post-hoc interpretable methods [3].

The experiments are performed on two datasets: the well-known MNIST[1], and the BDD100K dataset[2]. We empirically compare the selected models, analyzing the explanations given (Fig. 1 shows sample predictions of BagNet), examining also some of the components of these models from different perspectives (the confusion matrices of Fig. 2 show the pairwise separability of the resulting prototype vectors of ProtoPNet, demonstrating the importance of the separation cost).

[1]Modified National Institute of Standards and Technology database, http://yann.lecun.com/exdb/mnist/

[2]Berkeley DeepDrive dataset, https://bdd-data.berkeley.edu/



(a) (b)

Figure 2. Analysis of pairwise separability of the prototypes obtained for ProtoPNet on the MNIST dataset, using linear SVMs: (a) 20 epochs with $Sep = -0.008$, avg. accuracy: 0.9788, (b) 80 epochs with $Sep = 0$, avg. accuracy: 0.8477.

References

[1] W. Brendel, M. Bethge: *Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet*, arXiv preprint arXiv:1904.00760, 2019.

[2] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, C. Rudin: *This looks like that: deep learning for interpretable image recognition*, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 8930–8941, doi: 10.5555/3454287.3455088.

[3] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal: *Explaining explanations: An overview of interpretability of machine learning*, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 80–89, doi: 10.1109/DSAA.2018.00018.

[4] B. Goodman, S. Flaxman: *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"*, AI Magazine 38.3 (2017), pp. 50–57, doi: 10.1609/aimag.v38i3.2741.

[5] O. Li, H. Liu, C. Chen, C. Rudin: *Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions*, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018, pp. 3530–3537, doi: 10.1609/aaai.v32i1.11771.

[6] W. Samek, T. Wiegand, K.-R. Müller: *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*, arXiv preprint arXiv:1708.08296, 2017.

[7] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu: *dEFEND: Explainable Fake News Detection*, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 395–405, doi: 10.1145/3292500.3330935.