SDSC

ETH zürich

Building self-explained classification models

Michele Volpi, Nathanaël Perraudin, Radhakrishna Achanta, and Fernando Pérez-Cruz

January 2022

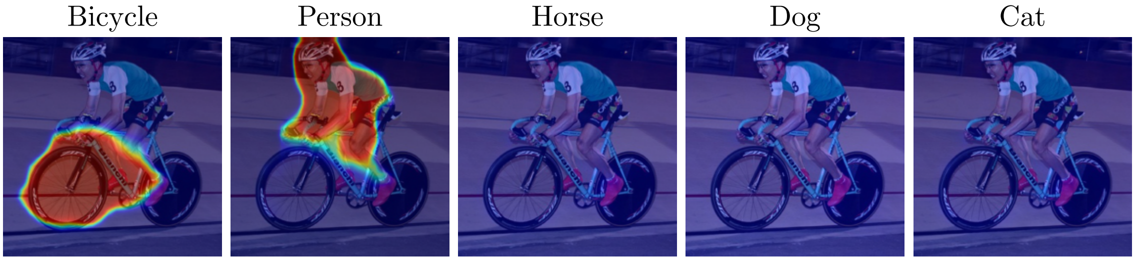

Figure 1: Visual comparison of **per-class** attributions provided by an *Explainer* method.

1 Summary

The immense success of deep learning in creating powerful models comes at the cost of poor interpretability. This limits the usefulness of deep learning in several applications where interpretability is crucial. In this project, we aim to improve interpretability [1], [2] by separating the useful from the useless information in a data sample, a task often referred as *attribution*.

Practically, we focus on computer vision image classification problems. The “useful” parts of the image to predict a given object class are separated from the rest using a mask. This mask can then be used as an explanation for the classification decision of the neural network. As summarized in Figure 2, the task of the student will be to develop an architecture that provides an explanation (the mask) additionally to the solution. The project consists in trying the ideal solution and make the necessary adjustments in order for it to work. We will explore, get inspired and compare with traditional solutions such as GradCam [3]–[6] or attention-based object localization schemes. An example of such attribution is visible in Fig. 1, where an attribution method is trained to mask parts of the images that explain the classification.

This research project is *challenging* but we already have working architecture for a similar problem.

2 Project description

The project consists of the following steps. i) The problem will first be studied from a theoretical perspective. The different losses will be derived. ii) A literature search and survey from similar methods [7], [8] will be conducted by the student. iii) A synthetic dataset will be constructed such that background and object are known in advance. iv) The technique will first be tried

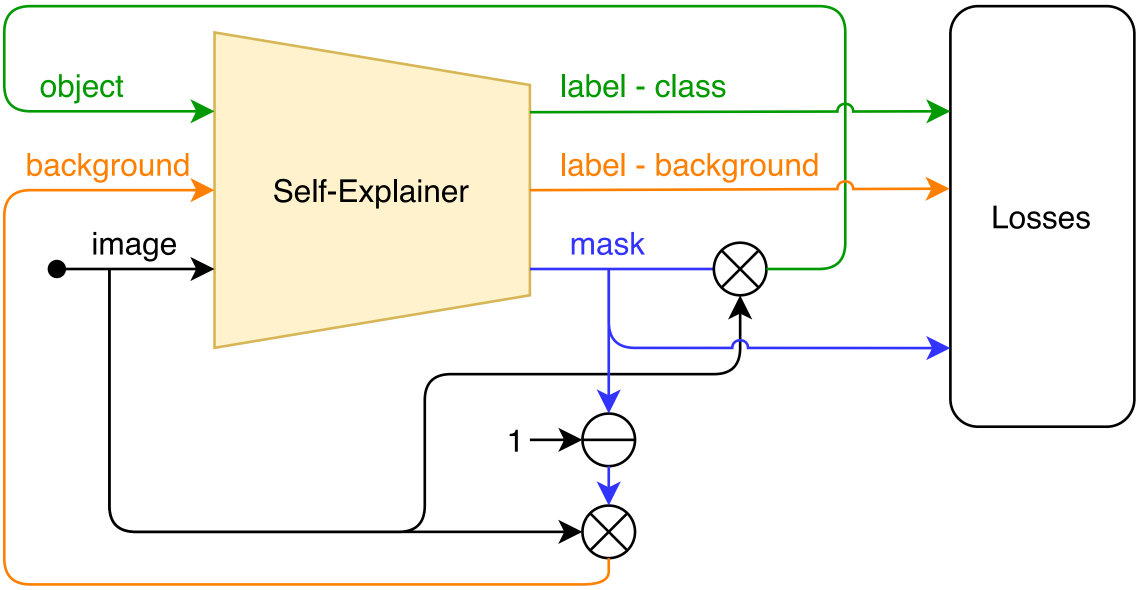

Figure 2: *General intuition of the architecture. Given an image, the Self-Explainer (typically a U-Net) constructs masks that isolates the useful information (the objects in the given class) from the background. Note that these masks can be easily be transformed into labels. Using the masks, the objects and the background can be isolated and fed separately into the Self-Explainer providing us with labels for each objects and for the background.*

on the synthetic dataset and then on datasets of increasing complexity (e.g. MNIST, CIFAR, ImageNet, MS-COCO, etc.).

3 Additional information

- **Difficulty of the project:** Challenging (but we guide you all along)!
- **Dataset:** MS-COCO, CUB200, Pascal VOC, MNIST, CIFAR, ImageNet
- **What will you learn?** Deep learning (CNNs), good scientific research practices, literature research, PyTorch.
- **Requirements:** Machine Learning fundamentals, linear algebra, computer vision fundamentals, good Python skills, experience with git, *motivation*.
- **Supervisors:**
  - Michele Volpi (michele.volpi@sdsc.ethz.ch)
  - Nathanaël Perraudin (nathanael.perraudin@sdsc.ethz.ch)
  - Radhakrishna Achanta (radhakrishna.achanta@epfl.ch)
  - Fernando Pérez-Cruz (fernando.perezcruz@sdsc.ethz.ch).

Bibliography

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Computing Surveys*, vol. 51, no. 5, Aug. 2018.

[2] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, pp. 206–215, May 2019.

[3] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-cam: Why did you say that?” *Neural Information Processing Workshop*, 2016.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *The IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[5] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018.

[6] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, *Score-cam: Score-weighted visual explanations for convolutional neural networks*, 2019.

[7] R. Fong, M. Patrick, and A. Vedaldi, “Understanding deep networks via extremal perturbations and smooth masks,” in *The IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019.

[8] V. Petsiuk, A. Das, and K. Saenko, “Rise: Randomized input sampling for explanation of black-box models,” in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.