Sanity checks for patch visualisation in prototype-based image classification - Supplementary material -

Romain Xu-Darme[1,2], Georges Quénot[2], Zakaria Chihani[1], Marie-Christine Rousset[2]

[1] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France
{romain.xu-darme, zakaria.chihani(at)cea.fr}
{georges.quenot, marie-christine.rousset(at)imag.fr}

1. Training prototype-based classifiers

In this work, we use the code provided by the authors of ProtoPNet [1] and ProtoTree [2] in order to train classifiers on the CUB-200-2011 dataset [3] (CUB) and the StanfordCars dataset [4]. For ProtoPNet, we use the default training parameters provided by the authors and trained the models during 50 epochs. For ProtoTree, we also use the default parameters and trained the models during 100 epochs. Table 1 presents the final accuracy of the models used in our experiments.

Table 1. **Accuracy of the self-explaining models** used in this work. CUB-c denotes the cropped CUB-200-2011 dataset.

Model	Backbone	Dataset	Accuracy
ProtoPNet	VGG19	CUB-c	75.1%
	ResNet50	CUB-c	72.5%
		CARS	71.4%
ProtoTree	ResNet50	CUB	83.1%
		CARS	83.2%

2. More prototype visualisation with Smoothgrads and PRP

Fig. 1 illustrates how using more faithful visualisation methods, such as PRP or Smoothgrads, rather than upsampling can improve the trust that the user can have in the model. In these examples, the upsampling strategy shows image patches focused on the background and gives a false sense of bias in the model, while PRP and Smoothgrads - which provide more faithful saliency maps - are focusing on elements of the bird.

3. Area under the Deletion Curve

In this section, we illustrate the evolution of the similarity ratio when incrementally removing the most important pixels of the image according to the saliency maps proposed by the different visualisation methods under study. As shown in the example of Fig. 2, removing pixels according to upsampling has little to no effect on the similarity score, suggesting that the explanation is incorrect. On the contrary, when removing only 1% of the image according to Smoothgrads, the similarity score drops to roughly 15% of its original value, suggesting that the explanation focuses on actual regions of interest for the model. The same result is achieved when removing only 0.3% of the image according PRP, indicating an even more precise explanation. Moreover, reaching a similarity ratio lower than 10% indicates that the explanation method has successfully identified the most relevant pixels of the image patch and gives an indication on the effective size of the image patch.

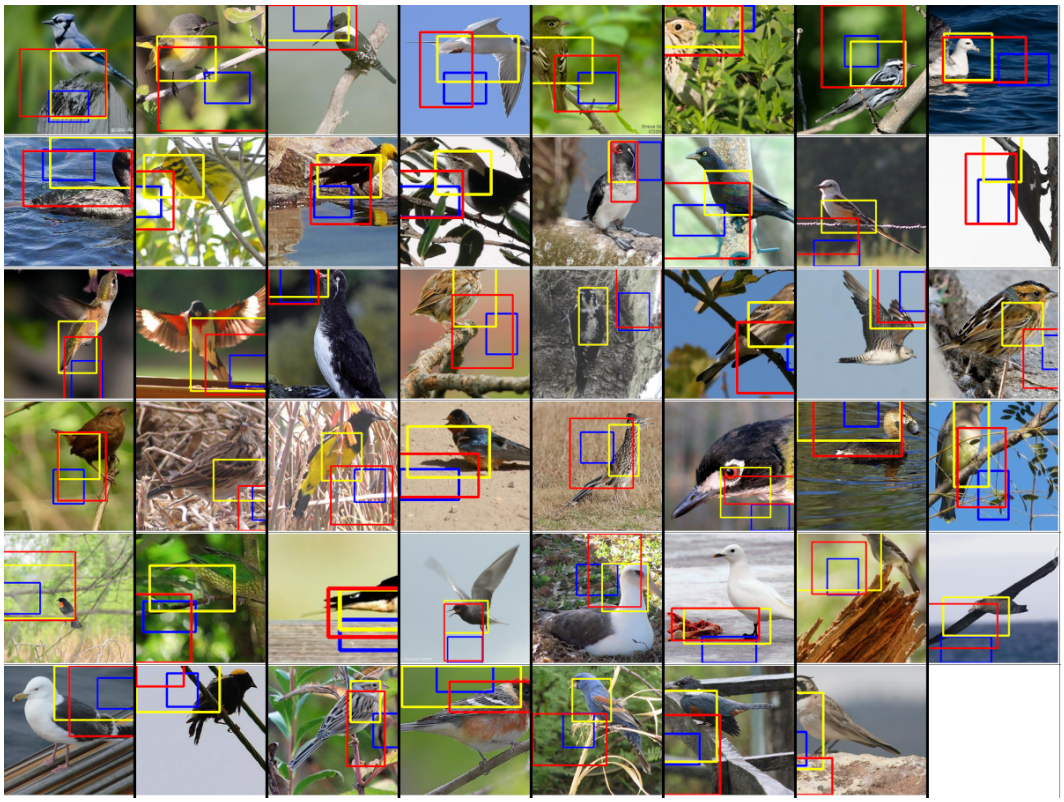


Figure 1. More examples of visualization of prototypes from a ProtoTree trained on CUB-200-2011 using upsampling with cubic interpolation (blue), Smoothgrads (red) or PRP (yellow).

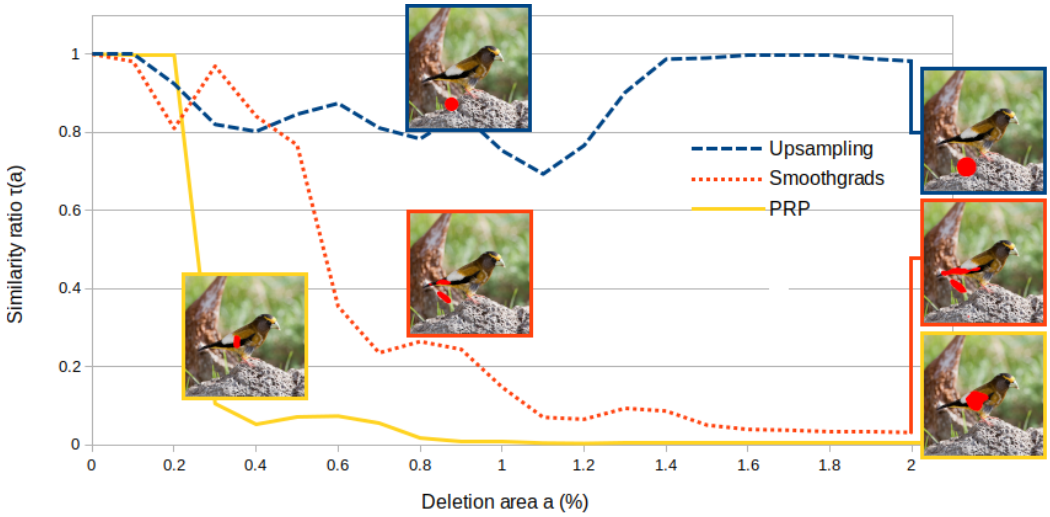


Figure 2. Evolution of the similarity ratio when incrementally removing the most important pixels according to the ProtoPNet/ProtoTree method (upsampling, in blue), Smoothgrads (red), and PRP (yellow). Best viewed in colour.

4. Distribution of similarity ratio v. deletion area on ProtoTree visualisation

In addition to the results presented in the paper focusing on the average similarity ratio v. deletion area, in this section we study the *distribution* of similarity ratios for a given deletion area (here 0.5%, 1%, 1.5% and 2%). As shown in Fig. 3, we notice a "sandglass" effect on the distribution of similarity ratios for ProtoTree prototypes: for low deletion areas ($\leq 1\%$), the

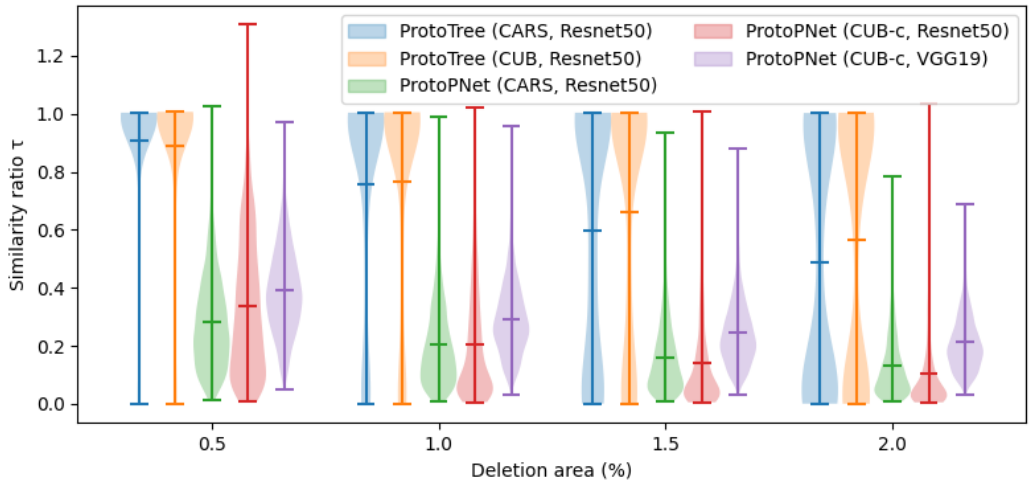
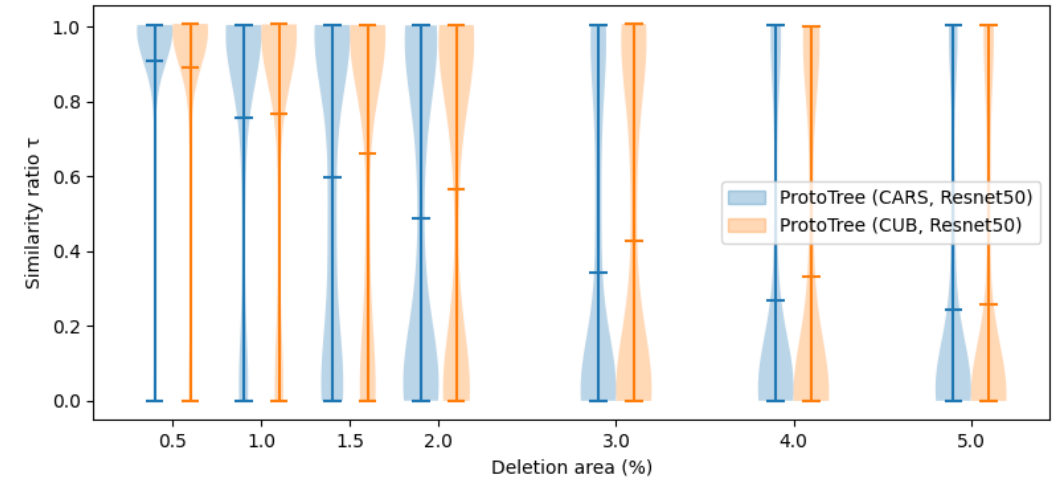
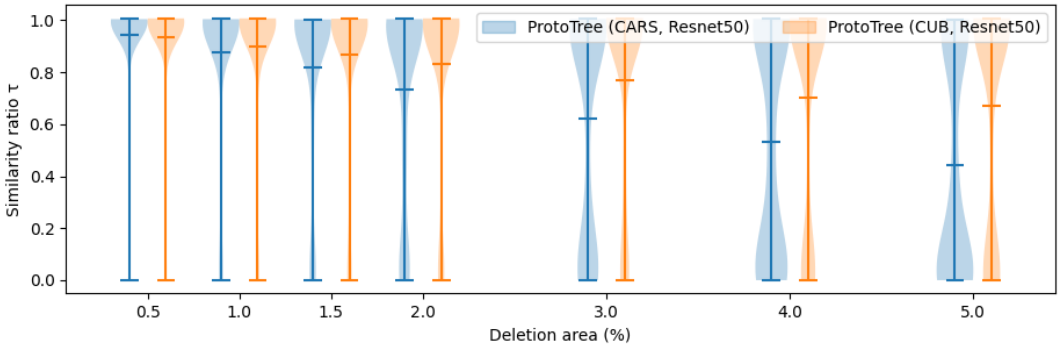


Figure 3. Distributions of similarity ratios v. percentage of deletion area when visualising prototypes using PRP.



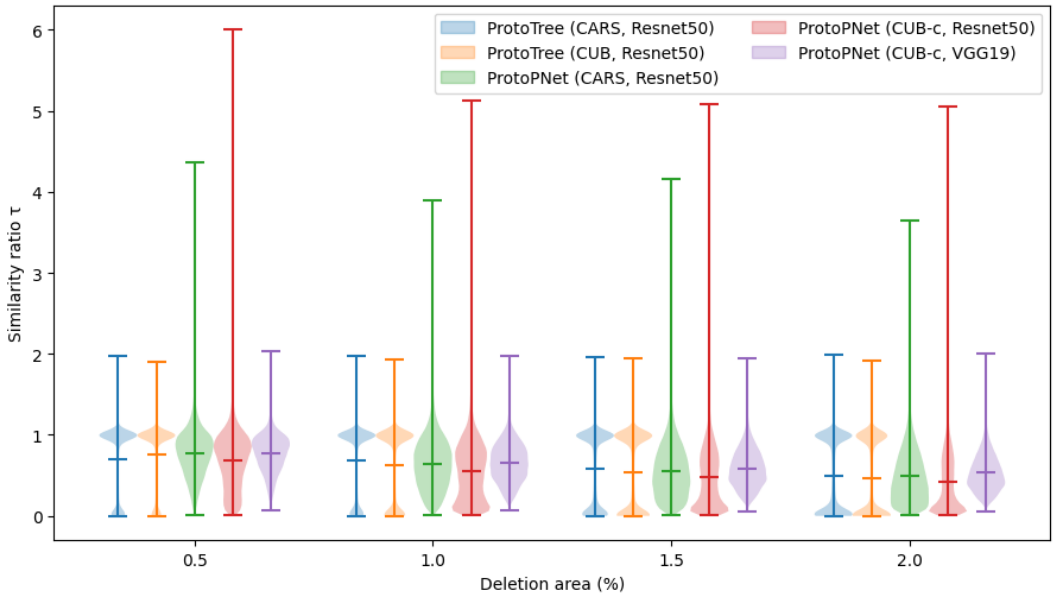
(a) Using PRP.



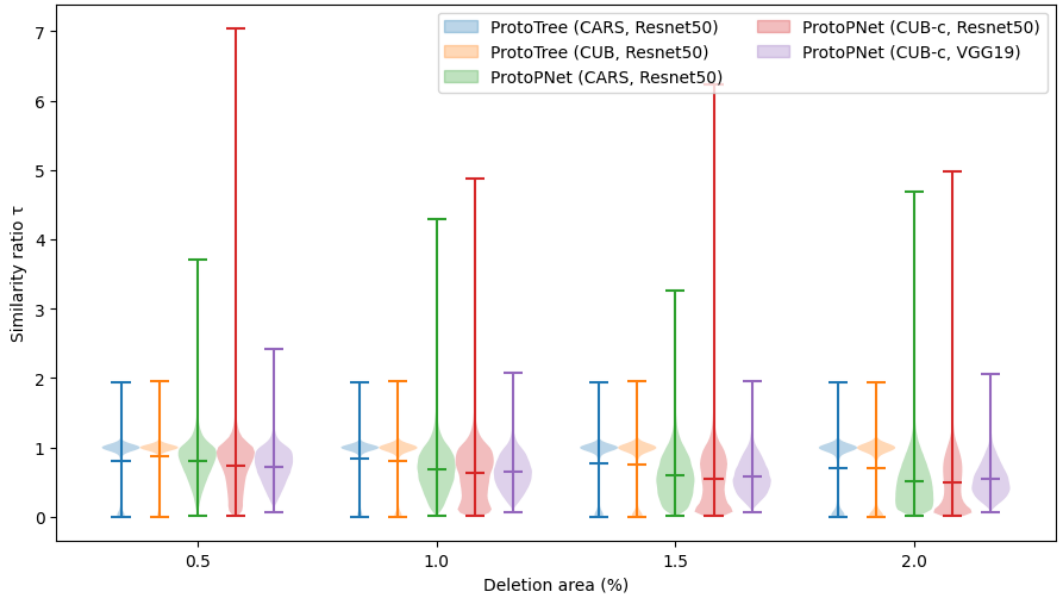
(b) Using Smoothgrads.

Figure 4. Distributions of similarity ratios v. percentage of deletion area when visualising ProtoTree prototypes

similarity ratio for all prototypes is close to 1. Then, from 1.5% up to 4-5% (Fig.4a), the distribution of similarity ratios slowly shifts towards 0. This suggests that the drop in similarity does occurs uniformly for all prototypes, but rather in a "continuous" manner, *i.e.* that ProtoTree prototypes have a wider range of size for their corresponding effective receptive fields than



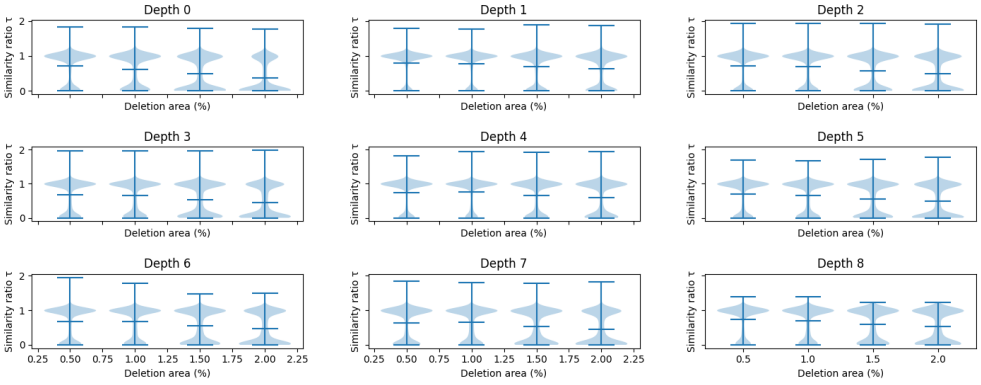
(a) Using PRP



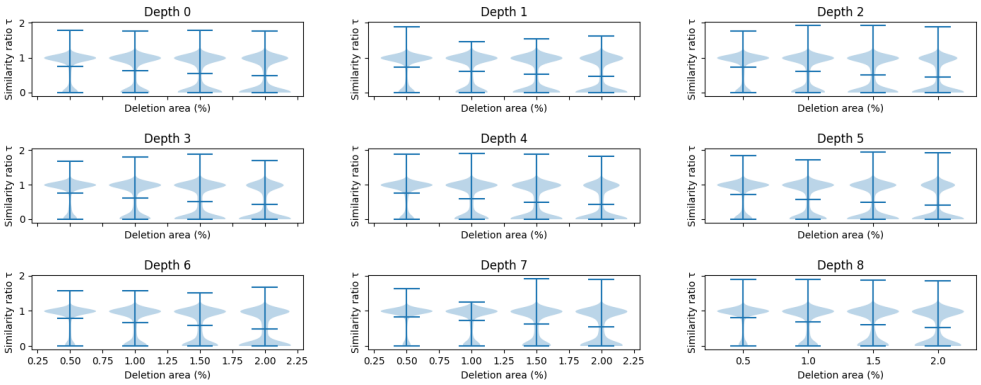
(b) Using Smoothgrads

Figure 5. Distributions of similarity ratios v. percentage of deletion area when visualising test patches during inference.

ProtoPNet prototypes. Moreover, as shown in Fig. 4b, the sandglass effect is also present when using Smoothgrads and when visualising image patches during inference (Fig. 5). Moreover, as shown in Fig. 6, this effect is seemingly uncorrelated to the depth of the prototype inside of the decision tree. This suggests that ProtoTree does not necessarily focus on finer - and smaller - details in the last stages of the decision process.



(a) On CARS



(b) On CUB

Figure 6. Distributions of similarity ratios v. percentage of deletion area when visualising ProtoTree test patches using PRP during inference. Results are sorted by depth inside of the decision tree.

References

[1] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. *This* looks like *That*: Deep learning for interpretable image recognition. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, page 8930–8941, 2019. 1

[2] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14928–14938, 2021. 1

[3] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, 2010. 1

[4] L. Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A Large-Scale Car Dataset for Fine-Grained Categorization and Verification. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3973–3981, 2015. 1