Deep Prototypical-Parts Ease Morphological Kidney Stone Identification and are Competitively Robust to Photometric Perturbations

Daniel Flores-Araiza[1] Francisco Lopez-Tiro[1] Jonathan El-Beze[2] Jacques Hubert[2]
Miguel Gonzalez-Mendoza[1] Gilberto Ochoa-Ruiz[1] Christian Daul[3]

[1]Tecnologico de Monterrey, School of Engineering and Sciences, Mexico
[2]CHU Nancy, Service d'urologie de Brabois, Nancy, France
[3]CRAN UMR 7039, Université de Lorraine and CNRS, Nancy, France

S1. Supplementary Material

S1.1. Performance under visual perturbations

Accuracy evaluated on the mixed test dataset, for the different ProtoPNet backbones and the number of PPs learned was obtained to identify the level of loss in performance of the models learning PPs. With this evidence for the kidney stone classification task made available, future users will have a reference for which visual characteristics PPs models tend to be more susceptible to and to which degree. It is found that hue perturbation on the input images is on average the perturbation with the most dramatic loss in performance for models learning PPs training with the Kidney Stones datasets. In contrast, this type of architecture tends to be robust against moderated changes in the brightness and saturation of the input images. With this evidence, a call for precaution when dealing with assumed small perturbation on the input images [17] is implemented.

Also, the performance of the traditional CNN used as feature extractors were explored under the same perturbations, these results are shown in the first four rows under the OOD performance metrics in Table 3 for each of the CNN backbones explored and their performance for IID test data and OOD is summarized in Fig. S3a. These same evaluations are shown for the different backbones of PPs models trained in Fig. S3b. It is observed PPs models present a lower standard deviation when compared to their CNN counterparts for IID evaluations, per this behavior for each of the different backbones used by the models, as seen in Fig. S3c.

S1.2. Prototypes projection details

The Projection of Prototypical parts (PPs) is the intermediate step A_2, mentioned in Section 4, performed in the training process that allows visualization of the learned PPs. In this step, each prototype p_j is assigned the value of its nearest latent training patch $f(x)_l$ from all the images of the same class k initially assigned to p_j. Therefore, this distance is $d_{j,l,k} = \|p_{j,k} - f(x)_{l,k}\|_2$. In this way, each abstract PPs learned conceptually equates to one training image patch. Allowing the faithful visualization of the learned PPs, use to generate the visual explanations as well as the final output classification given by the model. Mathematically, for the projection prototype p_j of class k, i.e., $p_j \in P_k$, we perform the following update:

$$p_{j,k} \leftarrow \underset{f(x)_{l,k} \in S_{train}}{\arg\min} \|p_{j,k} - f(x)_{l,k}\|_2 \tag{1}$$

where $S_{train} = \{f(x)_{l,k} : f(x)_{l,k} \in f(x_k)$ for all training images S_{train}, $(x, y) : y \in k\}$.

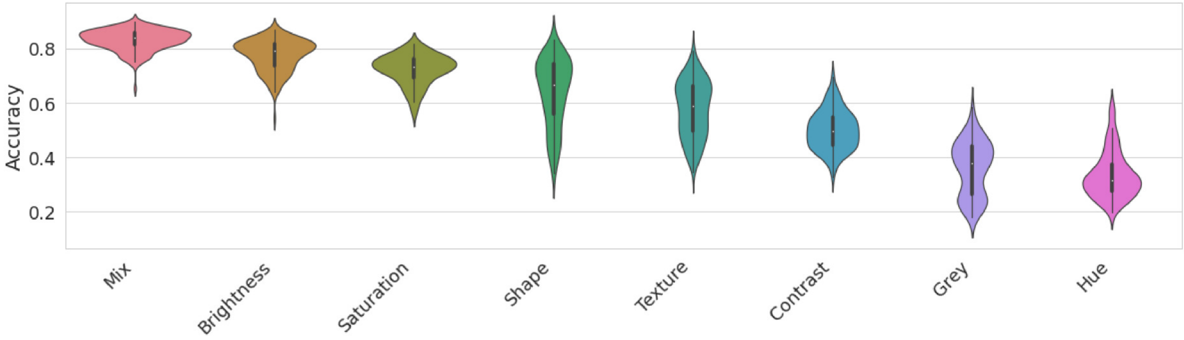


Figure S1. Accuracy evaluated on the test dataset, averaging the different ProtoPNet backbones and the number of Prototypical parts learned. It is found that hue perturbation on the input images is on average the perturbation with the most dramatic loss in performance for ProtoPNet models training with the Kidney Stones dataset. In contrast, the ProtoPNet architecture tends to be robust against moderated changes in the brightness and saturation of the input images.

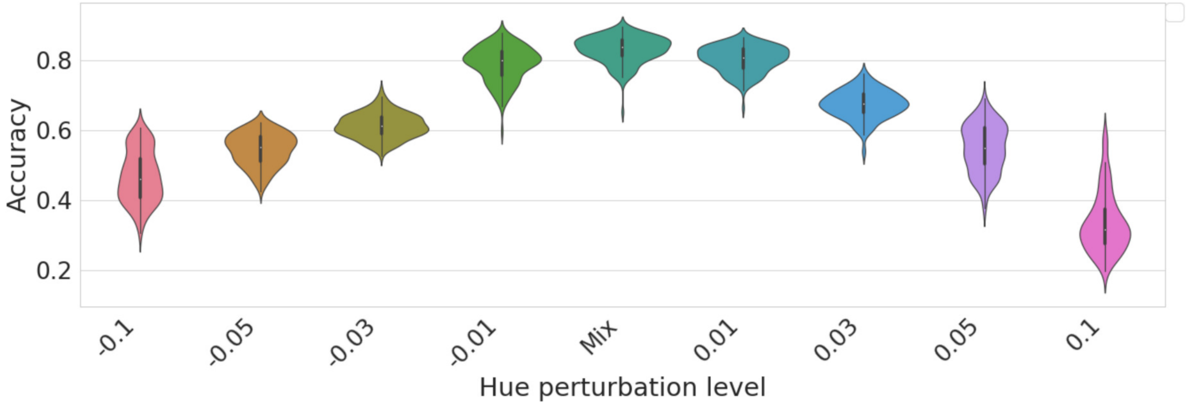
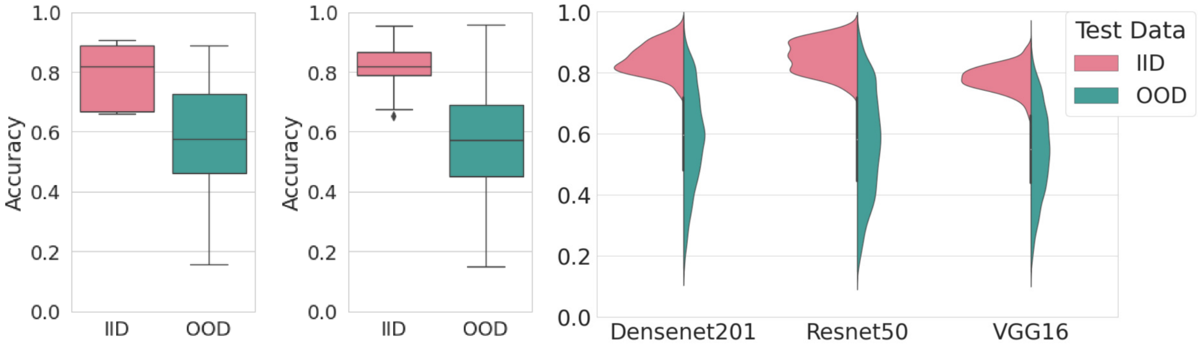


Figure S2. Accuracy evaluated on mixed test dataset under modifications of the images hue channel, with values between -0.1 and 0.1. These evaluations reflect the level of robustness ProtoPNet achieves with a DenseNet201 CNN backbone with 3 part-prototypes per class, under the perturbation that most affect the model.



(a) CNNs (b) ProtoPNets (c) ProtoPNets per CNN backbone

Figure S3. (a) Accuracy of CNN models evaluated separated on Independent and Identically Distributed (IID) test data and Out Of Distribution (OOD) tests. (b) Accuracy of ProtoPNet models evaluated on IDD and OOD test data. (c) Same as in ”b” for each of the three backbones used.