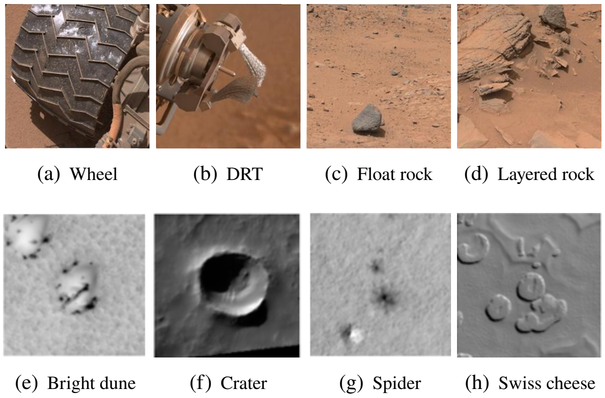
## PDS IMAGE CONTENT-BASED SEARCH WITH EXPLAINABLE MACHINE LEARNING

S. Lu[1], B. Vasu[1,2], E. Dunkel[1], K. Grimes[1], and M. McAuley[1]. [1]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA (you.lu@jpl.nasa.gov), [2]Oregon State University, Corvallis, OR 97331, USA.

**Introduction:** The NASA Planetary Data System (PDS) maintains archives of data collected by NASA missions that explore our solar system. The PDS Cartography and Imaging Sciences Node (Imaging Node) provides access to millions of images of planets, moons, comets, and other bodies. With the large and continually growing volume of data, there is a need for tools that enable users to quickly search for images of interest. Each image product archived at the PDS Imaging Node is described by a rich set of searchable metadata properties such as local season and the time it was collected. However, users often wish to search on the content of the image to zero in on those images most relevant to their use cases. Manually searching through millions of images is infeasible. In this abstract, we will summarize the machine learning (ML) classifiers created for the content-based search, and we will also introduce the recent advances in training explainable ML classifiers.

**Content-Based Search:** The image content-based search capability is deployed at the PDS Image Atlas (the Atlas) website[1]. The PDS Imaging Node currently maintains and operates the image content-based search capabilities for five missions. These content-based search capabilities are enabled using ML classifiers. In this work, we will summarize the techniques we used to create MSLNet and HiRISENet classifiers.

MSLNet is a hybrid of two Convolution Neural Network (CNN) classifiers[2] created for images collected using the Mast Camera (Mastcam) and Mars Hand Lens Imager (MAHLI) instruments onboard the Mars Science Laboratory Curiosity rover, and HiRISENet is a CNN classifier for images collected using the High Resolution Imaging Science Experiment (HiRISE) instrument onboard the Mars Reconnaissance Orbiter. Both MSLNet and HiRISENet were trained using transfer learning techniques [1] with data sets[3,4] we published on Zenodo. Example images we used to train and evaluate MSLNet and HiRISENet are shown in Figure 1.



(a) Wheel (b) DRT (c) Float rock (d) Layered rock

(e) Bright dune (f) Crater (g) Spider (h) Swiss cheese

**Figure 1:** Example images used to train and evaluate MSLNet (first row) and HiRISENet (second row).

The MSLNet and HiRISENet classifiers use a confidence threshold of 0.9 to determine which classification results will be shown to users of the Atlas. To ensure that the classifiers' self-reported posterior probabilities are well calibrated, we employed Platt scaling [2] to adjust the output for each class. Platt scaling selects a temperature $T$ and bias $b$ to transform the logits $Z_i$ output by the classifier for image $x_i$ prior to the conversion of $Z_i$ into a posterior probability $p = \frac{1}{1+e^{-Z_{i,k}/T_k+b_k}}$, where $Z_{i,k}$ is the logit for image $i$ and class $k$. The parameters $T$ and $b$ are optimized using L-BFGS algorithm with respect to the cross entropy loss on the validation set[5].

**Explainable Classifiers:** CNN classifiers are black-box models, which means their internal reasoning processes are not interpretable. Despite the performance metrics (e.g., accuracy, precision, recall, F-score, etc.) one usually computes for evaluation, it is still possible for a CNN classifier to learn unexpected features that potentially could lead to biased conclusions. To overcome the stated challenges, we explored the prototypical part network (ProtoPNet) architecture proposed in [3]. Moreover, we integrated the ProtoPNet architecture with the MSLNet and HiRISENet to provide human-understandable explanations. We chose ProtoPNet because (1) it helps with ensuring that the features the classifiers learned are meaningful, and (2) the explanations can intuitively be presented to the users of the Atlas.

The ProtoPNet architecture uses a standard deep network for feature extraction and introduces a prototypical layer that learns a pre-defined number of prototypes that best describe the class. The fully connected layer after the prototypical layer represents the contribution of each learned prototype to the final classification. The learned

[1] https://pds-imaging.jpl.nasa.gov/search

[2] The MSLNet v1 classifier focuses on rover hardware (e.g., wheel, calibration target, etc.), and the MSLNet v2 classifier focuses on science targets (e.g., float rock, layered rock, etc.)

[3] HiRISE data: https://doi.org/10.5281/zenodo.4002935

[4] MSL data: https://doi.org/10.5281/zenodo.4033453

[5] To train and evaluate classifiers, we split the MSLNet and HiRISENet data sets into training, validation, and test sets

**Table 1:** Experimental results of MSLNet and HiRISENet classifiers (best performance scores are in bold). Note that “Acc (0.9)” in the title means the accuracy score computed with a 0.9 confidence threshold, and “P-” in the Classifiers column indicates the classifier was created with a prototypical layer.

| | Test Set Performance | | |
|---|---|---|---|
| **Classifiers** | Acc | Acc (0.9) | Abst Rate |
| MSLNet (AlexNet) | 74.5% | **87.2%** | 36.2% |
| MSLNet (VGG19) | **81.3%** | 85.6% | **12.0%** |
| MSLNet (ResNet18) | 79.5% | 86.4% | 15.5% |
| P-MSLNet (VGG19) | 75.2% | 82.5% | 19.8% |
| P-MSLNet (ResNet18) | 74.8% | 83.0% | 24.5% |
| HiRISE (AlexNet) | 92.8% | 94.4% | 7.1% |
| HiRISE (VGG19) | 92.6% | 94.9% | 6.7% |
| HiRISE (ResNet18) | **93.5%** | **95.0%** | 3.5% |
| P-HiRISE (VGG19) | 91.7% | 93.2% | **3.1%** |
| P-HiRISE (ResNet18) | 91.5% | 92.9% | 4.9% |

prototypes contain information about the classifiers' internal reasoning process and can be projected onto regions of the test image. Once the regions are found, they can be visualized as bounding boxes by using a threshold at 95% of maximum similarity. The similarity score represents the strength of the prototype match, while the weights of the fully connected layer represent its contribution to a class during training.

**Results:** We explored different options for training MSLNet and HiRISENet classifiers. Firstly, we trained the classifiers with VGG19 [4] and ResNet18 [5] backbone architectures. Secondly, we trained the classifiers with and without prototypical layers. The classifiers' performance results are summarized in Table 1. The classifiers with the AlexNet [6] backbone architecture are currently deployed on the Atlas. For MSLNet, the classifier trained with AlexNet backbone architecture yields the best threshold accuracy score of 87.2%, but its abstention rate of 36.2% is the worst compared to other classifiers. For HiRISENet, the classifier trained with ResNet18 backbone architecture yields the best threshold accuracy score of 95.0%. The classifiers with the prototypical layers perform slightly worse than the classifiers without the prototypical layers, which is expected as the addition of prototypical layers constrains the information passed down to the final layers.

An example visualization of the explanation generated with the “P-MSLNet (VGG19)” classifier is shown in Figure 2. Firstly, one can observe the visual similarity between the prototypes for the test and train images in columns (c) and (d). Secondly, the prototypes with positive weight scores (i.e., the first two rows) should increase the predicted probability that the test image belongs to the “Wheel” class; the prototypes with negative weight scores (i.e., the last three rows) should decrease the predicted probability that the test image belongs to the “Wheel” class.

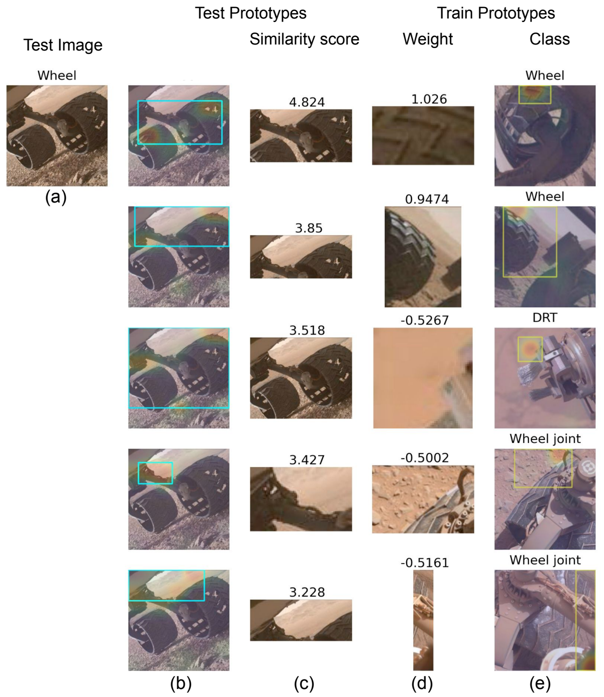


**Figure 2:** Example visualization of the explanation generated with the“P-MSLNet (VGG19)” classifier. Column (a) is the test image, and column (b) is the same image overlayed with a heatmap showing regions most activated by the prototype learned during training, followed by column (c) showing a cropped version of the heatmap after the threshold with the similarity score. Column (e) shows the training images overlayed with regions obtained after prototypes projection on the training set and column (d) shows the cropped regions of heatmaps from Column (e)

**Conclusion & Future Work:** In this abstract, we summarized the techniques used to train and evaluate MSLNet and HiRISENet, and we also demonstrated the use of prototypical layers to train explainable classifiers. In terms of future work, we will (1) investigate an ensemble approach that could potentially improve the performance of the explainable classifiers; (2) calibrate the posterior probabilities of the explainable classifiers; and (3) deploy the explainable classifiers to the next generation Atlas website that is being developed now.

**Acknowledgments:** Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

**References:** [1] Wagstaff, K. *et al.* (2021) *IAAI.* [2] Platt, J. (1999) *Advances in Large Margin Classifiers*. [3] Chen, C. *et al.* (2019) *NeurIPS*. [4] Simonyan, K. *et al.* (2015) *ICLR*. [5] He, K. *et al.* (2016) *CVPR*. [6] Krizhevsky, A. *et al.* (2012) *NeurIPS*.