IMPROVED CONTENT-BASED IMAGE CLASSIFIERS FOR THE PDS IMAGE ATLAS. S. Lu[1], K. L. Wagstaff[1], J. Cai[2], G. Doran[1], K. Grimes[1], J. Lee[3], L. Mandrake[1], Y. Yue[2], [1]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109-8099, USA (you.lu@jpl.caltech.edu), [2]California Institute of Technology, Pasadena, CA, 91125, USA, [3]Columbia University, New York, NY, 10027, USA.

Introduction: The Planetary Data System (PDS) Imaging Node hosts millions of images acquired from the planet Mars. Missions such as the Mars Science Laboratory (MSL) and Mars Reconnaissance Orbiter (MRO) are actively collecting new images to enrich our understanding of Mars. These new images are delivered to the PDS Imaging Node periodically. All of these images are served for public access by the PDS Image Atlas (Atlas).

With the constantly growing image volume, connecting scientists, engineers, and the general public to images of interest has become a challenge. Images delivered to the PDS Imaging Node are required to contain metadata in PDS standards. The metadata contains descriptive information regarding when and how the images are processed and transferred to the Earth. However, users of the Atlas are often interested in finding images based on content, and the content-based information is not included in the metadata, and must be extracted through content analysis.

To enable content analysis for efficiently finding images of interest, we proposed a solution that uses a deep convolutional neural network (CNN) [1]. Deep CNN classifiers were shown to achieve high performance on a variety of computer vision challenges in 2012 [2]. Training deep CNN classifiers from scratch usually requires thousands to millions of labeled images. To reduce this labeling effort, we utilized transfer learning to adapt a network previously trained using Earth images (CaffeNet [3]) for use with Mars orbital and surface images. The initial versions of the two classifiers, HiRISENet and MSLNet, were deployed on the Atlas for public use in 2017.

Data Sets: The data set used for HiRISENet consists of 10,433 Mars landmark images cropped from 180 HiRISE map-projected images. We augmented this data set using rotation, flipping, and brightness adjustment methods to obtain a total of 73,031 landmark images covering eight classes[1]. The data set used for MSLNet consists of 6,691 MSL images collected by the Mastcam left eye, Mastcam right eye, and MAHLI (Mars Hand Lens Imager)[2]. The labels span 24 classes of engineering interest (e.g., parts of the rover). Examples from both data sets are shown in Figure 1.

Recent Improvements: Over the past year, we have employed several methods to improve both the accuracy of the classifiers and the reliability of their confidence values.

[1]HiRISENet data: https://zenodo.org/record/2538136
[2]MSLNet data: https://zenodo.org/record/1049137

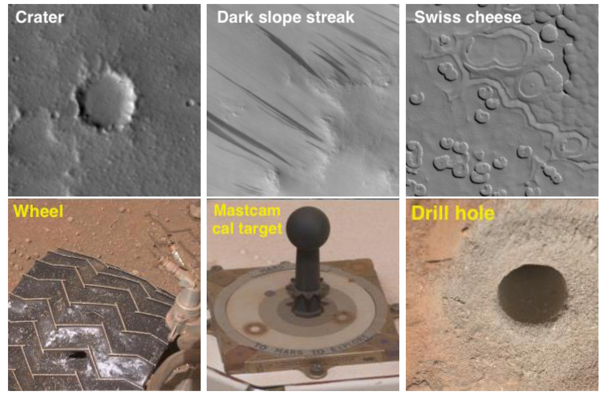

Figure 1: Example images from the HiRISENet (top) and MSLNet (bottom) data sets.

First, we modified our fine-tuning methodology. The original CaffeNet model was trained for 310,000 iterations on 1.2 million ImageNet images from 1000 classes. To fine-tune this model for HiRISE or MSL data, we previously used a learning rate multiplier of 1 for layers 1 to 7 and 10 for the final layer (fc8) [1]. This allowed only small changes to the previously trained layers and larger adaptation for the output (classification) layer. While investigating possible improvements to the MSL classifier, we found that setting the multiplier to 0 for layers 1 to 4 ("freezing" them) enabled better generalization.

Next, we analyzed each classifier's errors in the training and validation sets to determine the most common errors and thereby develop a strategy for further improving classifier accuracy. We created an interactive browser-based Error Analysis Tool to enable fast review of classification errors in the training and validation data sets. For the HiRISE classifier, we found that the most common classifier errors arose from (1) inconsistent crater labels (e.g., some craters were missed; some degraded craters were incorrectly labeled by humans as "other") and (2) failing to detect very faint, very small, or off-center dark slope streaks. Both observations led us to review and update the labels for these classes in the entire data set. For the MSL classifier, we found that the "drill hole" class (Figure 1, lower right) was the most common culprit among the validation set errors. On further inspection, we discovered that all 36 drill hole images in the training set were views of the same drill hole, thus limiting the classifier's ability to recognize new drill holes. This inspired us to obtain more labeled drill hole images from the training set timespan (MSL sols 3 to 181).

Finally, we employed classifier calibration [4] to improve the reliability of self-reported posterior probabilities. We found that the most effective method for HiRISENet was temperature scaling, which identifies a temperature parameter T that is applied to the logit values z_i output by the classifier for item x_i prior to the conversion of z_i into a posterior probability $p = \max_k \sigma(z_{i,k}/T)$, where $\sigma(\cdot)$ is the softmax function and $z_{i,k}$ is the logit for item i and class k. Calibration of MSLNet is in progress.

Results: HiRISENet performance was generally higher than MSLNet performance. The HiRISE data set is more balanced and representative than the MSL data set, and the total number of classes (8 versus 24) is lower, so the problem may be inherently easier. The results are shown in Table 1. Augmenting the HiRISE data set improved training and generalization performance. Modifying the fine-tuning ("FT") had little impact for HiRISE but yielded major improvements in performance for MSL.

Table 1: Classification accuracy for HiRISENet and MSLNet.

Classifier	Train	Val	Test
HiRISE	88.3%	88.6%	84.3%
HiRISE+aug.	98.1%	**91.8%**	90.0%
HiRISE+aug.+FT	**98.2%**	91.2%	**90.2%**
MSL	98.7%	72.8%	66.7%
MSL+FT	**99.8%**	**84.0%**	**76.9%**

Classifier calibration using temperature scaling does not change the predictions (or overall accuracy); instead, it impacts accuracy at a given confidence cutoff (e.g., HiRISENet uses a 90% confidence threshold to decide which results are shown to Atlas users) as shown in Table 2. Classifier reliability is measured by the Expected Calibration Error (ECE), a weighted sum of the difference between reported confidence and empirical accuracy across multiple confidence bins [4]. HiRISENet ECE improved (lower is better) while test accuracy did not change significantly. Reliability diagrams are a visual representation of classifier calibration [4]. The reliability diagram for HiRISENet after temperature scaling is shown in Figure 2. The resulting confidence values align closely with the (optimal) diagonal.

Table 2: ECE and classification accuracy at 90% confidence.

Classifier	ECE	90% Test acc.
HiRISE+aug	0.056	93.45%
HiRISE+aug+calib.	**0.036**	**93.46%**

PDS Image Atlas: The Atlas[3] is an interactive browser-based interface that uses the Apache Solr indexing system. Apache Solr implements Apache Lucene syntax which powers the search and navigation features of the Atlas. The image content-based search capability is enabled in the Atlas via a two-step process: (1) the classification results of HiRISENet and MSLNet are indexed by Apache Solr, and (2) the indices are made searchable in the Atlas. In addition to content-based search, the Atlas supports a variety of functionalities that allow users to easily and intuitively navigate millions of archived images.

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

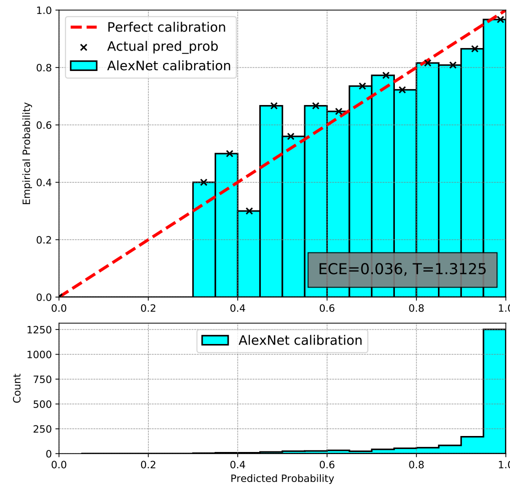

Figure 2: Reliability diagram (top) and posterior probability distribution (bottom) for HiRISENet after calibration.

Future Work: Collecting high-quality labels is a major challenge for any machine learning system. We are evaluating active learning techniques to further reduce the labeling cost and enable easy expansion to images collected by other missions [5]. The class distribution of the data sets used to fine-tune HiRISENet and MSLNet are severely imbalanced and change over time, so we are also exploring techniques to handle domain shift.

Acknowledgments: The authors wish to thank Sue LaVoie and Paul Ramirez from PDS Imaging Node for their continuous support of this work. 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. L. *et al.* (2018) *IAAI*. [2] Krizhevsky, A. *et al.* (2012) *NIPS*. [3] Jia, Y. *et al.* (2014) *arXiv:1408.5093*. [4] Guo, C. *et al.* (2017) *ICML*. [5] Settles, B. (2010) *Computer Sciences Technical Report 1648*.