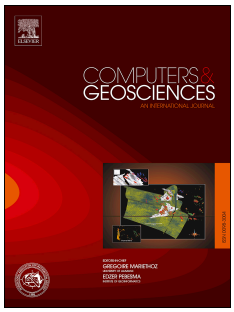
Accepted Manuscript



Context-dependent image quality assessment of JPEG compressed Mars Science Laboratory Mastcam images using convolutional neural networks

Hannah R. Kerner, James F. Bell, III, Heni Ben Amor

PII: S0098-3004(17)30968-8

DOI: 10.1016/j.cageo.2018.06.001

Reference: CAGEO 4141

To appear in: *Computers and Geosciences*

Received Date: 13 September 2017

Revised Date: 1 May 2018

Accepted Date: 1 June 2018

Please cite this article as: Kerner, H.R., Bell III., , J.F., Ben Amor, H., Context-dependent image quality assessment of JPEG compressed Mars Science Laboratory Mastcam images using convolutional neural networks, *Computers and Geosciences* (2018), doi: 10.1016/j.cageo.2018.06.001.

Context-Dependent Image Quality Assessment of JPEG Compressed Mars Science Laboratory Mastcam Images using Convolutional Neural Networks

Hannah R. Kerner[a,*], James F. Bell III[a], Heni Ben Amor[b]

[a]*School of Earth and Space Exploration, Arizona State University, 781 E. Terrace Mall, Tempe, AZ 85287*
[b]*School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, 699 South Mill Ave, Tempe, AZ 85281*

Abstract

The Mastcam color imaging system on the Mars Science Laboratory *Curiosity* rover acquires images that are often JPEG compressed before being downlinked to Earth. Depending on the context of the observation, this compression can result in image artifacts that might introduce problems in the scientific interpretation of the data and might require the image to be retransmitted losslessly. We propose to streamline the tedious process of manually analyzing images using *context-dependent* image quality assessment, a process wherein the context and intent behind the image observation determine the acceptable image quality threshold. We propose a neural network solution for estimating the probability that a Mastcam user would find the quality of a compressed image acceptable for science analysis. We also propose an automatic labeling method that avoids the need for domain experts to label thousands of training examples. We performed multiple experiments to evaluate the ability of our model to assess context-dependent image quality, the efficiency a user might gain when incorporating our model, and the uncertainty of the model given different types of input

☆H. Kerner performed this research effort including developing all algorithms and code, conducting all experiments, and writing this manuscript. J. F. Bell provided guidance on Mars Science Laboratory mission procedures and the context-dependent quality assessment of Mastcam images, and edited this manuscript. H. Ben Amor provided guidance on the design and evaluation of the proposed machine learning solutions, and edited this manuscript.

*Corresponding author
Email address: hkerner@asu.edu (Hannah R. Kerner)

images. We compare our approach to the state of the art in no-reference image quality assessment. Our model correlates well with the perceptions of scientists assessing context-dependent image quality and could result in significant time savings when included in the current Mastcam image review process.

Keywords: deep learning, machine learning, planetary science, image quality

1. Introduction

The Mastcam color imaging system on the Mars Science Laboratory *Curiosity* rover acquires images within Gale crater for a variety of geologic and atmospheric studies (*e.g.*, Malin *et al.*, 2017, Bell *et al.*, 2017, Grotzinger *et al.*, 2012). Images are often JPEG compressed onboard the rover before being downlinked to Earth. While critical for transmitting images on a low-bandwidth connection, this compression style can result in small image artifacts most noticeable as anomalous brightness or color changes within or near 8×8-pixel JPEG compression block boundaries (Fig. 1). In high-frequency detail regions of some images, for example in regions showing fine layering or lamination in sedimentary rocks, the image must be retransmitted losslessly (*i.e.*, without lossy JPEG compression) to avoid introducing difficulties in the scientific interpretation of the data. The process of identifying which images have been adversely affected by compression artifacts is performed manually by the Mastcam science team. As of sol 1928, Mastcam acquired 87,885 images and 18,800 ($\sim$21%) of these were retransmitted losslessly. This process requires a significant time commitment from human experts and consumes critical portions of the available downlink data volume.

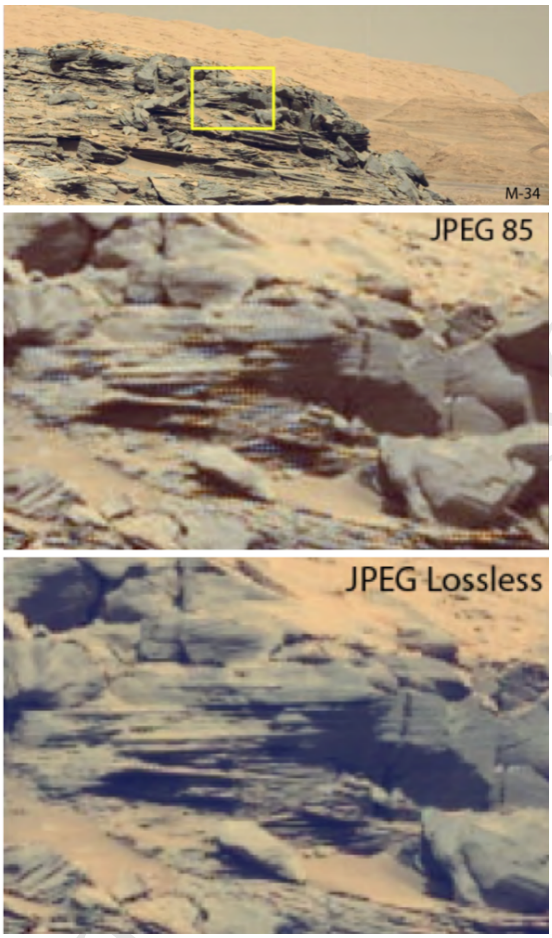


Figure 1: Mastcam M-34 image of finely-layered outcrop rocks acquired on *Curiosity* sol 1155, sequence mcam05219. The middle inset shows a zoomed-in view of some of the layers in the image original downlinked with JPEG compression factor 85. The bottom image is an example of the same scene without compression. (Adapted from Bell *et al.*, 2017)

In this work, we aim to facilitate the scientific image review process using *context-dependent* image quality assessment. We define context-dependent image quality assessment as a process wherein the context and intent behind the image observation determine acceptable image quality thresholds. We propose to automatically identify images where quality might be problematic using a two-part machine learning solution. Our proposed solution relies on: 1) a lo-

gistic regression model that maps compression level and joint entropy between an uncompressed and compressed image to the *image utility*, defined as the probability that a scientist would accept the quality of the compressed image; and 2) a convolutional neural network (CNN) that learns to predict the image utility given only the pixel information in the compressed image. Our solution can characterize the perceived quality of an entire image or small image patches. To evaluate this methodology, we perform an experiment to compare the time and effort expended by a Mastcam scientist when identifying images to retransmit. We show experimentally that, when assisted by our proposed method, a Mastcam investigator could significantly reduce the time required to review images. We also present a user study that surveys Mastcam data users to assess the correlation between assessments by our model and perceptions of context-dependent image quality by scientists.

This paper is organized as follows: Section 2 details previous work related to context-dependent image quality assessment. In Section 3 we describe our source dataset. In Section 4 we present our method for automatically labeling examples for training. In Section 5 we present our CNN model for assessing context-dependent image quality. Section 6 details the experiments and results for evaluating our proposed method and Section 7 discusses the contribution of these results. Finally, Section 8 summarizes our conclusions and proposes directions for future work.

2. Related work

2.1. No-reference image quality assessment

Previous works have proposed methods for no-reference image quality assessment (NR-IQA), also called blind image quality assessment, which quantifies and predicts the perceived quality of a distorted (*e.g.*, JPEG-compressed) image without access to a reference image (*e.g.*, the uncompressed image). However, existing approaches rely on one or all of the following: (1) off-the-shelf or hand-crafted features, (2) a definition of image quality that is independent of

the subject in the image, or (3) benchmark datasets, such as the LIVE Image Quality Assessment Database (Sheikh *et al.*), for demonstrating performance. These aspects inhibit their use in real-world problems where the context of the image *and* the level of distortion are important for quality assessment.

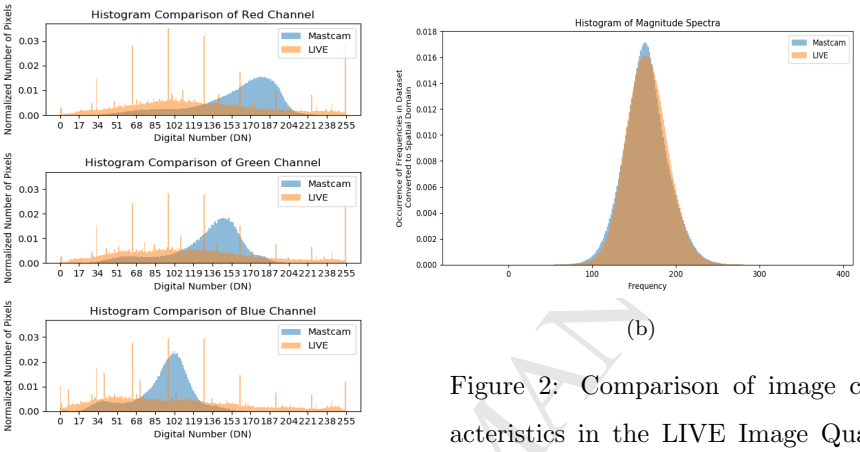
Most NR-IQA approaches manually design and extract features that are discriminant for quality degradations resulting from compression or other distortions. Successful approaches such as BRISQUE (Mittal *et al.*, 2012), DIIVINE (Moorthy *et al.*, 2011), BLIINDS-II (Saad *et al.*, 2012), Ghadiyaram *et al.*, 2014, and Hou *et al.*, 2015 commonly employ Natural Scene Statistics (NSS) for discriminative features to estimate quality as a measure of naturalness. Other successful approaches (Wang *et al.*, 2002, Li *et al.*, 2011, Chetouani *et al.*, 2015) use hand-crafted features based on statistical properties computed from the image.

Recent approaches for NR-IQA have demonstrated state-of-the-art results using automatically learned features for estimating image quality. Ye *et al.*, 2012 proposes an unsupervised feature learning technique based on codebook representations. Other recent work demonstrated the automatic feature learning capability of neural networks. Tang *et al.*, 2014 used a three-layer deep belief network to learn higher-level representations from pixels used as features in Gaussian Process regression to predict image quality scores. Bianco *et al.*, 2017 used a pre-trained CNN to automatically extract features describing generic image distortions for a support vector regressor predicting image quality scores. These works use two-step processes of 1) automatic feature extraction, followed by 2) regression using extracted features to predict image quality. Kang *et al.*, 2014 combines these steps into a single optimization procedure where features are extracted in a single convolutional layer and regression is performed in two fully-connected layers of a neural network. Additionally, Kang *et al.*, 2014 works on patches of input images to enable local image quality assessment. Of previous works on NR-IQA, this work is most similar to ours in that the authors propose an end-to-end CNN solution to assess image quality of local image patches that are combined for whole-image assessment.

The primary difference between previous work and our work is that in previous work, the image quality assessment is independent of the image subject. These works perform objective image quality assessment for generic distorted images, whereas our solution predicts the context-dependent image quality of JPEG-compressed images acquired for geologic study. In this application of image quality analysis to scientific images, quality is not an objective measure of feature distortion as in other work. It is a context-specific measure that represents the likelihood that artifacts introduced during compression will complicate the scientific analysis of the image. This is an important distinction because compressing a Mastcam image might significantly reduce the perceived quality of the image without affecting the scientific utility of the image. This might be the case if the observation was not intended for scientific analysis (but to monitor damage to the rover's wheels, *e.g.*) or scientific analysis of the image is not affected by the compression artifacts (since the scale of the target of analysis is much larger than the scale of compression artifacts).

The LIVE Image Quality Assessment Database (Sheikh *et al.*) is frequently used for training and assessing performance of IQA models (Wang *et al.*, 2002, Wang *et al.*, 2005, Wu *et al.*, 2013, Soundararajan *et al.*, 2012, Moorthy *et al.*, 2011, Chaofeng *et al.*, 2011, Saad *et al.*, 2012, Mittal *et al.*, 2012, Tang *et al.*, 2014, Kang *et al.*, 2015, Chetouani *et al.*, 2010, Hou *et al.*, 2015, Kang *et al.*, 2014). This database is relatively small (982 images for all distortions and only 233 for JPEG) so it cannot be used for training machine learning models like neural networks that require extensive training data. We performed an experiment to compare image characteristics of the LIVE (Earth) database with our Mastcam (Mars) dataset (Fig. 2). We found that the two datasets did not differ significantly in the frequency domain (Fig. 2b). However, we found that histograms of the red, green, and blue color channels differ significantly between the two datasets (Fig. 2a). The histograms of pixel value distributions in each channel of our Mastcam source dataset (described in Section 3) are Gaussian distributed with clear peaks. In contrast, the histograms of each channel of the LIVE dataset appear closer to a uniform distribution, which might be expected

for images of assorted everyday subjects on Earth. The spikes and gaps seen in the LIVE database histograms are due to compression artifacts in the images. We do not see these artifacts in the Mastcam histograms because these images have not been compressed yet.



(a) (b)

Figure 2: Comparison of image characteristics in the LIVE Image Quality Assessment Database (Sheikh *et al.*) of Earth images and our Mastcam dataset of Mars surface images. Comparing histograms of all images in each dataset after applying the Fast Fourier Transformation (b) shows that the two datasets have similar frequency distributions. However, histograms of pixel values in red, green, and blue channels across the datasets show very different color distributions; the Mastcam distribution is Gaussian and the LIVE distribution is nearly uniform. The spikes and gaps seen in the LIVE database histograms are a result of blocky compression artifacts in the images.

2.2. Reduced-reference image quality assessment

The automatic labeling approach we propose is most closely related to reduced-reference image quality assessment (RR-IQA). RR-IQA measures automatically quantify and predict the perceived quality of a distorted (*e.g.,* JPEG-compressed) image with partial access to a reference image. Wang *et al.*, 2005

measures image distortion using the KL-divergence between the marginal probability distributions of wavelet coefficients of the reference and distorted images. Soundararajan *et al.*, 2011 proposes an information theoretic framework that measures the distance between the reference image and the projection of the distorted image onto the space of natural images. Wu *et al.*, 2013 proposes a method informed by the human visual system that separately computes and evaluates the orderly portion of the image (the primary visual information) and the disorderly portion (the residual uncertainty). We propose a similar approach to Wang *et al.*, 2005 that uses joint entropy, a measure of uncertainty between two distributions, and the compression level to estimate the perceived information lost during compression. To our knowledge, ours is the first work that uses RR-IQA to automatically label a dataset used for NR-IQA, thus reducing the requirement for large hand-labeled datasets.

3. Source Dataset

The images for our training and test datasets are sourced from the NASA PDS-released Mastcam database of uncompressed images called RecoveredProducts that were previously retransmitted losslessly. We use RGB images collected between sols (Martian days) 121-1087 using both the M-100 (medium angle, right "eye") and M-34 (narrow angle, left "eye") imagers for training data and those collected between sols 1537-1672 for test data. Dividing the dataset into train and test sets by date of acquisition rather than a percentage split better represents how our model will be used in practice, *i.e.*, predicting the utility of images collected over time as the rover traverses new geologic regions. The resulting source dataset contains 6,911 images for training and 1,719 images for testing.

Full-resolution images from this source dataset are sliced into patches of 160×160 pixels for training. The specific size of 160×160 pixels was chosen to reveal a large enough region of the full observation to infer the geologic context, but a small enough region to "zoom in" on compression artifacts (on the order

of 8×8 pixels) and reduce the time required for training (Fig. 3). Training on patches also enables local assessment of image quality. This is important because the frequency of detail can vary significantly across a single image. Using image patches also significantly increases the size of our training dataset.

For the CNN training and testing data, we use a stride size of 200 pixels which yields 36 patches of size 160×160 pixels in each 1200×1344-pixel image. The resulting dataset contains 310,680 examples taken from 8,630 source images. For the automatic labeling model, we use the same source images but a different stride size (163 pixels) than was used for generating the CNN dataset (Fig. 4). This ensures that duplicate image patches are only possible when the product of the number of slices and the stride size reaches a common multiple of the two stride sizes. Since the maximum value for this product in our dataset is well below the least common multiple of 200 and 163 (32,600), we can guarantee that there will be no overlap between image patches used for training the logistic regression and CNN model.

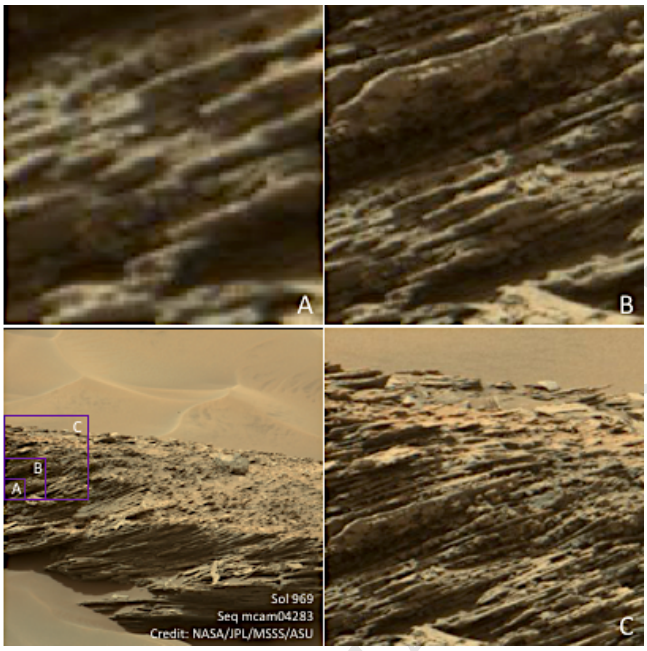


Figure 3: Images A, B, and C show 80 × 80-, 160 × 160-, and 320 × 320-pixel patches of a full-resolution Mastcam image respectively. The 160 × 160-pixel patch size best maximizes the obviousness of artifacts while still allowing some geologic context to be inferred, which is important for labeling the images.

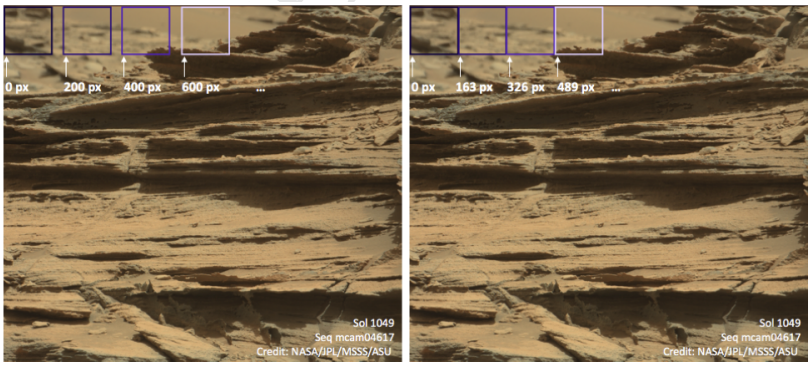


Figure 4: Images used in training the CNN were sliced from Mastcam images in our base training set with a stride size of 200 pixels, while images for the logistic regression automatic labeling model were sliced with a stride size of 163 pixels. This ensures that even though the same base Mastcam images might be used for training both models, the same image slices (which are the inputs to each model) would not be used for training both models.

4. Automatic labeling of training data

4.1. Perception of scientific image quality

A user of Mastcam data determines whether a downlinked JPEG-compressed image should be retransmitted losslessly based on both the scientific context of the image and the perceived level of distortion resulting from compression. A scientist might accept more distortion in an image where the intent is to understand the context of a study area (Fig. 5c) or to study low-frequency morphologies like sand dunes or boulders, the general shapes of which are not severely distorted by small-scale compression artifacts (Fig. 5d). Distortion can also be more acceptable in observations likely intended for engineering purposes, for example to check the general health of the rover's wheels or other subsystems (Fig. 5e). The level of distortion a scientist is willing to accept can vary in other images depending on how the compression artifacts affect the scientific interpretation of the image. For example, a scientist might accept low image quality for an observation where finer details are distorted by artifacts, but that distortion does not affect the scientific interpretation (Fig. 5a). In images containing very high-frequency features such as fine layering (lamination) or bedding in rocks (Fig. 5b), for example where scientists might wish to analyze properties of the layers such as frequency and spacing, most scientists require high image quality.

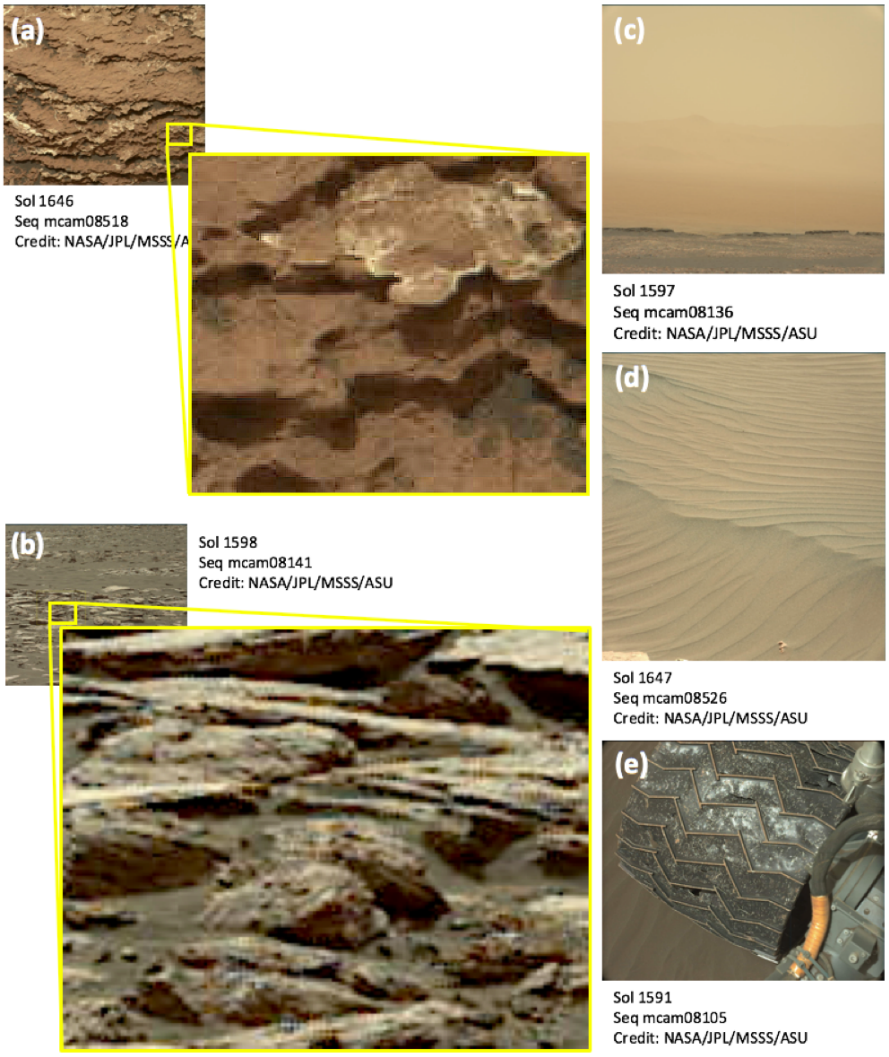


Figure 5: A scientist using the Mastcam dataset might have different requirements for image quality depending on the context of the observation. If the objective is to understand the context of a study area (c) or to study low-frequency morphologies like sand dunes or boulders (d), a higher level of compression might be acceptable. Distortion can also be more acceptable to scientists in observations likely intended for engineering purposes, for example to check the health of the rover's wheels or other subsystems (e). There might be some cases where distortion from compression is apparent but it does not affect the scientific interpretation of the image contents (a). In images containing very high-frequency features such as fine layering (lamination) or bedding in rocks (b), for example where scientists might wish to analyze properties of the layers such as frequency and spacing, scientists might demand high image quality.

4.2. Proposed approach for automatic labeling

Analyzing tens of thousands of images for training a CNN to label Mastcam images is prohibitively time consuming for scientists. A human-labeled dataset would require extensive participation from multiple domain experts in order to account for varying scientific interests in use of the Mastcam dataset. To reduce the effort required to label our training dataset, we propose an automatic labeling system that requires relatively few examples to be labeled and approximates the varying interests of scientists who use Mastcam data.

To create training data for automatic labeling, we randomly selected images from the source dataset described in Section 3 and compressed the selected images using a random quality between 75 and 95. Based on inputs from Mastcam experts with a variety of interests, we labeled images as "accept" if the quality of the image was acceptable given the context of the observation or "retransmit" if the scientific utility of the image might be compromised by compression artifacts. Since negative examples (labeled "retransmit") are much less common in the dataset than positive examples, images were manually labeled until 21 negative examples were identified. These were complemented by 21 positive examples for a total training set size of 42 images. We fit a logistic regression classifier using compression level and joint entropy between the compressed and uncompressed patch as features to predict the label a human would apply to an image. Joint entropy is a measure of uncertainty between two distributions and has been used in image processing to represent the difference between a pair of images (*e.g.*, Maes *et al.* (1997)).

We compute the joint entropy between the uncompressed and compressed versions of an image by first computing a joint histogram of pixel values between the two images, normalizing this histogram to yield a joint probability distribution, then computing the entropy of the joint probabilities (Korn and Korn (2000)):

Algorithm 1 Compute the joint entropy between two images

Input: $X_1, X_2 \in \mathbb{R}^N$ where N is the number of red, green, and blue pixels in an image

Output: Joint entropy between two images

function JOINTENTROPY(X_1, X_2)

initialize HIST $\in \mathbb{R}^{256\times256}$ where $\forall i, j \mid 0 \leq i \leq 256$ and $0 \leq j \leq 256$, HIST$[i][j] = 0$

for $0 \leq i < N$ **do**

HIST$[X_1(i)][X_2(i)] =$ HIST$[X_1(i)][X_2(i)] + 1$

end for

return $-\frac{1}{N}\sum_i \sum_j \text{HIST}[i][j] \log \text{HIST}[i][j]$

end function

In the next section, we describe how this classifier is used to automatically generate training data on the fly for training a CNN to predict the perceived quality of Mastcam images without a reference (lossless) image.

5. Convolutional neural network for predicting scientific utility

When compressed Mastcam images are downlinked from the rover, the science team must assess the context-dependent quality, which we term the scientific utility, of the image without a lossless version of the image to use as a reference. Without this reference image, we cannot use the automatic labeling system described in Section 3. To predict which Mastcam images contain distortions that might complicate scientific analysis, we propose a CNN to automatically learn the features for assessing scientific utility directly from a compressed Mastcam image.

We create a batch for training by compressing an image from our source dataset with a random quality between 75 and 95, then generating 36 patches of 160×160 pixels with a stride size of 200 pixels.

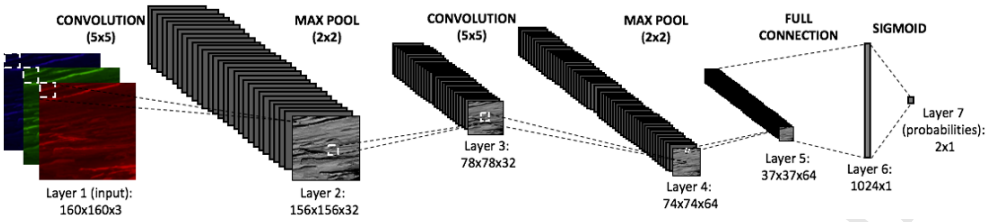


Figure 6: The convolutional neural network architecture.

Our CNN utilizes a standard architecture built using Google's Tensor Flow library for programming neural networks (Abadi *et al.*, 2015). The input to the network is the 160×160 pixel image patch, segmented into red, green, and blue color channels. The network (Fig. 6) contains two convolutional layers with 5×5-pixel kernels, each followed by a 2×2-pixel max pooling layer. The feature maps computed in the convolutional layers are input in the next layer to a fully connected layer of neurons. The neurons in both the convolutional layers and this fully connected layer utilize the rectified linear unit, or ReLU (Glorot *et al.*, 2011), to compute non-linear transformations of the data. This is followed by a "dropout layer" with keep probability 0.4 to reduce overfitting (Srivastava *et al.*, 2014), and finally the "readout" layer, which computes the log odds of scientific image quality acceptance. We apply the softmax function to convert these log odds into probabilities. The network learns to approximate the probability distribution of the labeled training data by minimizing the cross entropy between the modeled probability distribution and the examples seen during training. For this optimization, we use the cross-entropy with logits loss function and Adam Optimizer (Kingma and Ba, 2015) provided in the TensorFlow API. The scientific utility, or probability that a scientist would find the image quality acceptable given its context, is computed across the entire image by averaging the probabilities produced by the model for each of the slices across the image. The local utility estimation in each slice across an entire image is visualized in Fig. 7.

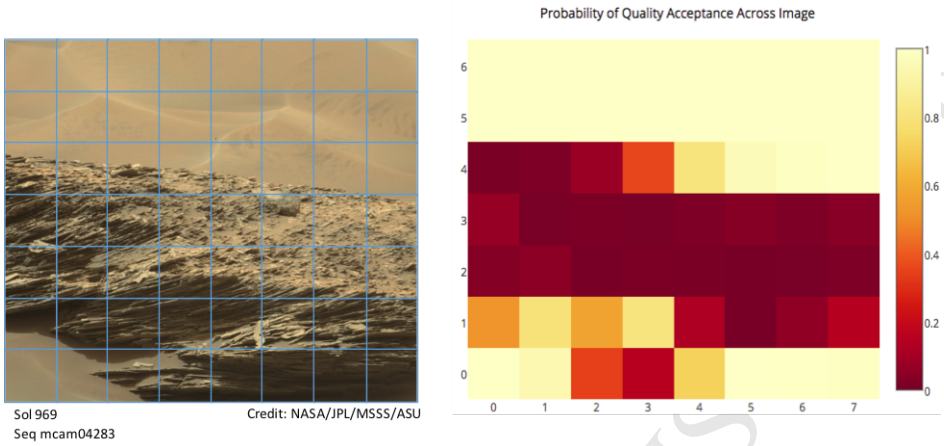


Figure 7: Using 160 × 160 pixel slices of a larger image as input to the CNN allows local perceived quality estimation to be performed. The image on the left shows a test image sliced into 160 × 160 pixel fragments with a stride size of 160 pixels. The heatmap on the right shows the distribution of acceptance probabilities across the entire image. These individual probabilities are averaged to estimate the probability that a user would find the image quality of an image acceptable or not (thus choosing to retransmit that image lossslessly or not).

6. Experiments

The proposed system for identifying images that might be requested lossless when evaluated by a Mastcam science team member is intended as an assistant rather than a replacement for scientists reviewing these images. In the current process, each of three team leaders (a principal investigator (PI) and two Deputy PIs) of the Mastcam imaging system must review hundreds of images every few months and cast a vote for each image to be either retransmitted losslessly or deleted from the Mastcam computer. In the future, a science team member could request from the proposed software a list of images where the perceived image quality is estimated to be below a certain percentage, for example 50%, or request a list of the images sorted by probability of quality acceptance. We present an experiment to compare the amount of time an investigator spends on this process with the current manual system and the estimated amount of time an investigator might spend on this process when assisted by our proposed

machine learning method. We also present results from a user study to assess the correlation between context-dependent image quality assessment by our model and by Mastcam data users. This study was approved by the ASU Institutional Review Board (ID STUDY00007622). We compare the performance of our CNN model for context-dependent image quality assessment to the state of the art in no-reference image quality assessment. We assessed the accuracy of our logistic regression classifier for automatic labeling on test data and compare the performance to other popular classifiers. Finally, we present an experiment to estimate the uncertainty of our model's predictions.

6.1. Accuracy of automatic labeling model

The test dataset for our logistic regression classifier is a set of 42 images selected randomly from the source dataset and manually labeled as "accept" or "retransmit". We did not balance the test set by class as we did for the training set. In Fig. 8, we plot feature values of the test dataset and the decision boundary determined by the parameters learned during training. Our logistic regression classifier achieves 83.3% accuracy on test data. We trained several popular classifiers and compare their performance on test data in Table 1. The highest accuracy is achieved by logistic regression and random forest, but we chose to use logistic regression because it is mathematically straightforward and more readily understood across disciplines. All test examples that were incorrectly classified were false negatives, meaning the classifier labeled some examples as "retransmit" that were accepted but did not label any examples as "accept" that were retransmitted. For scientists to adopt our method of identifying images that might need to be retransmitted, it is important for our automatic labeling classifier to have a low rate of false positives, or examples automatically classified as "accept" that should actually be retransmitted.

Classifier	Accuracy on Test Set
Logistic Regression	83.33%
Random Forest	83.33%
Neural Network	80.95%
Naive Bayes	80.95%
Support Vector Machine	76.19%

Table 1: Comparison of Popular Classifiers for Automatic Labeling Model

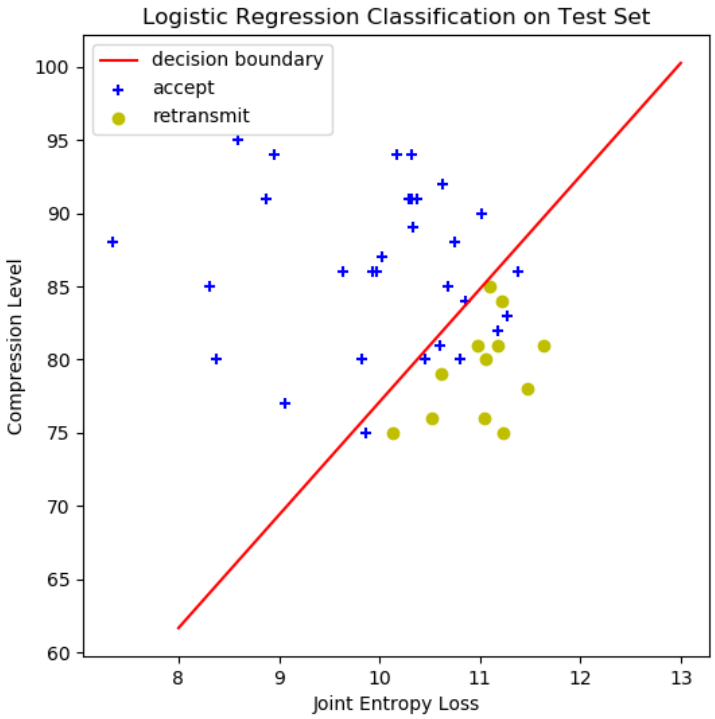


Figure 8: Our logistic regression classifier uses compression level and joint entropy between a compressed and uncompressed version of an image to predict the label a human would assign to the image. This classifier achieves 83.33% accuracy on test data.

6.2. Correlation of Model Predictions with Perceived Image Quality

To evaluate our CNN model's predictions for scientific image quality, we conducted a user study of Mastcam data users. Eleven users were shown the same set of 30 (160×160-pixel) image patches from compressed Mastcam images containing geologically diverse content in addition to the full-resolution image the patch came from. Each user was asked to rate on a linear scale from one to five the suitability of each image for the indicated intended analysis given their perception of the quality of the image. Scores below three were interpreted as recommendations to retransmit the image, scores above three as recommendations to accept the image, and scores equal to three as not sure. We computed the combined user recommendation as the sum of the "accept" responses and half of the "not sure" responses divided by the total number of participants for each image. Aggregate responses from participants and our CNN model estimates for each image are shown in Fig. 9.

There is significant variation in the responses from participants, even among those who study the same geologic processes and Mastcam image products. In general, our CNN's assessments of image quality given geologic context agree with the assessments made by participants in this user study. In all except four examples (Fig. 10), the model's prediction and the combined user prediction fall on the same side of the 50% decision boundary. While there is of course some error in the model, we also consider that some error might be due to differing interpretations of the questions or indicated intended analysis in the user study. For example, it is possible for a participant to have included factors of the image independent of compression artifacts (*e.g.,* framing) in their assessment of the suitability of an image for the intended analysis.

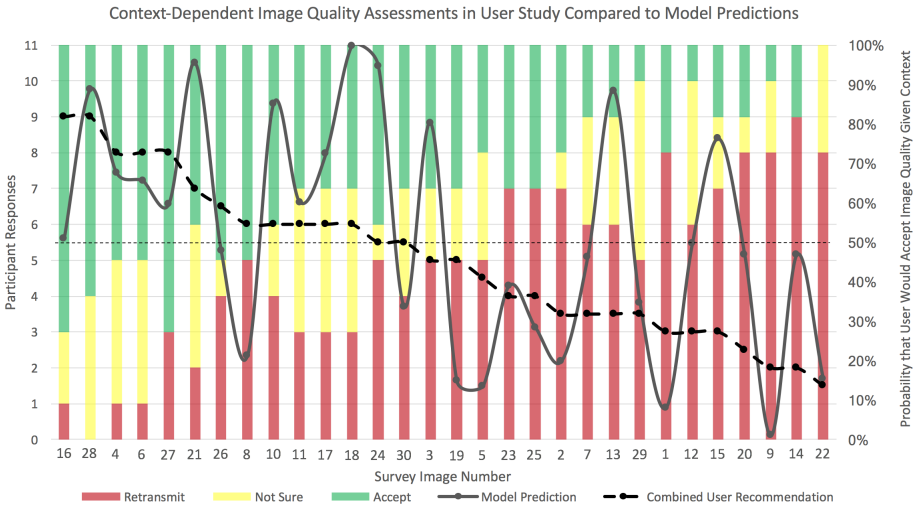


Figure 9: Model and combined participant estimates of the probability that a user would accept the quality of each image given its scientific context (right y-axis) superimposed on participant responses from user study for each image (left y-axis).

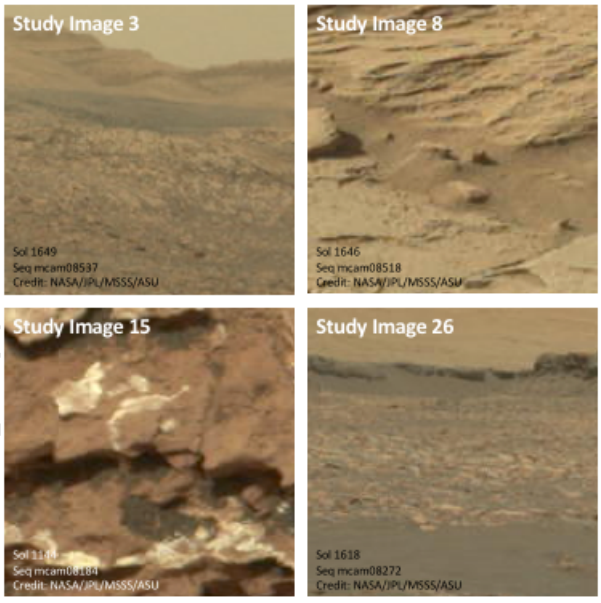


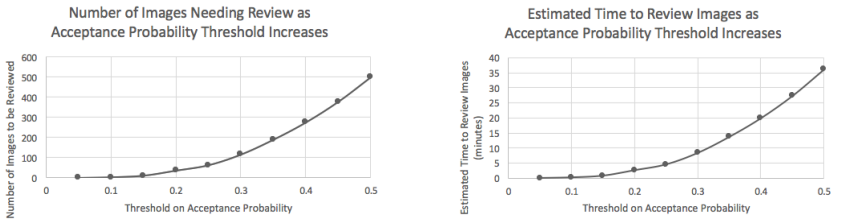
Figure 10: Images from user study where model recommendation and combined user recommendation disagreed.

6.3. Comparison of Manual and Assisted Methods

To estimate the amount of time an investigator spends reviewing images using the existing manual process, we observed a Mastcam Deputy PI conduct the review process for sols 1537-1672, which includes observations for 124 sols consisting of 1,719 compressed RGB images (not every sol has a downlink). The images are reviewed by sol, which might contain anywhere between a few to dozens of images depending on the observation plan for that sol. From this study, we estimated a lower bound of approximately one minute per sol that the investigator spends reviewing the images to mark them for deletion or lossless retransmission. Thus, for a typical review period of about 150 sols, an investigator can expect to spend at least two and a half hours reviewing images with the existing manual process. The results from this study were typical of previous image assessment activities conducted by this Deputy PI over the five year history of the MSL mission's Mastcam investigation.

Since we cannot truly measure the amount of time an investigator would spend in the image review process using the proposed method without modification to the existing internal mission operations software, we estimated this time by assuming the investigator would spend the usual time reviewing images down-selected by our model but negligible time reviewing those above the acceptability threshold set by the investigator. We ran our model on all images downlinked between sols 1537-1672 with quality acceptance threshold varying between 50% and 5%. We plotted these thresholds and the number of images our model classifies below that threshold in Fig. 11a. This plot shows that there is an approximately exponential increase in the number of images needing review as the acceptance threshold increases. We computed the expected time required to review images when assisted by our proposed method as a fraction of the time that would be spent reviewing all 1,719 images, and plotted this as a function of the acceptance probability threshold in Fig. 11b. Using a 50% threshold, the investigator would need to review about one third of the images when assisted by the proposed method than without. We estimate that an investigator would spend a maximum of $\sim$36 minutes to review images in the

studied sol range, compared to a lower bound of ∼124 minutes when unassisted.



(a) (b)

Figure 11: In the practical implementation of our proposed method, a science team member could request to review only a list of images with perceived image quality acceptance probability below some percentage, for example 50%, or request a list of the images sorted by probability of quality acceptance. The number of images below the selected threshold that the scientist might need to review is shown in the plot on the left. On the right, we plot the estimated time that might be spent reviewing the images (computed as a fraction of the time spent reviewing 1,719 images) as a function of the selected threshold.

6.4. Performance Compared to Related Work

As discussed in Section 2, the approach proposed in Kang *et al.*, 2014 for no-reference image quality assessment using a CNN is most similar to our approach for context-dependent image quality assessment. We trained our model and the implementation from Kang *et al.*, 2014 on JPEG-compressed Mastcam images obtained between sols 121-1087 and evaluated their performance on the same source dataset used in training but using different stride values (following the procedure in Kang *et al.*, 2014). We use the Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC) measures to evaluate model performance. Table 2 shows there is a good correlation between predictions by our model and labeled test data and that our model outperforms the Kang *et al.*, 2014 model on context-dependent image quality assessment. We note that Kang *et al.*, 2014 proposed a solution to the general no-reference image quality assessment problem, not *context-dependent* image quality assessment as we are proposing in this work. Despite this, we perform this comparison to the state of the art for completeness.

Model	LCC	SROCC
Kang *et al.*, 2014	0.101895	0.196932
Our CNN	0.546739	0.630751

Table 2: Comparison of our model proposed for context-dependent image quality assessment and model proposed by Kang *et al.*, 2014 for general no-reference image quality assessment

There are key differences between Kang *et al.*, 2014 and our approach that explain the difference in performance for context-dependent image quality assessment. The LIVE dataset for JPEG distortions is derived from 29 high-resolution RBG color images compressed with varying JPEG quality to produce a dataset of 233 images. LIVE provides a Difference Mean Opinion Score (DMOS) for each image. Scores range from 1 to 100 and are based on responses from observers about their perception of the quality of each image (Sheikh *et al.*). Kang *et al.*, 2014 uses the LIVE images and DMOS scores for training and testing. We modified the code provided by Kang *et al.*, 2014 to use Mastcam images from sols 121-1087 and the perceived image quality predictions generated by our logistic regression labeler (also in the range 1 to 100). Additionally, Kang *et al.*, 2014 pre-processes images by generating 32×32-pixel patches from the image and applying a local contrast normalization, which might discard information that is potentially useful for inferring contextual information about the image.

An important assumption in Kang *et al.*, 2014 is that the distortion in LIVE images is roughly homogeneous across the entire image, and thus the DMOS score for the entire image is used as the score for all patches in that image. While this may be a valid assumption for general purpose image quality assessment, it is not a valid assumption for context-dependent image quality assessment. As illustrated in Fig. 7, the perceived scientific quality of an image patch is highly dependent on the geologic features in that image. Using the same label for scientific image quality for all patches in a Mastcam image would make it difficult to learn a mapping between the pixels of an image patch and scientific

image quality in the entire image.

6.5. Evaluation of Model Uncertainty

Gal and Ghahramani, 2016 presented a method based on dropout (Srivastava *et al.*, 2014) for estimating the predictive uncertainty of a neural network. Gal and Ghahramani show that an arbitrarily deep, non-linear neural network with dropout applied before every weight layer is mathematically equivalent to an approximation of the probabilistic deep Gaussian Process. They also show that by performing many stochastic forward passes through a trained neural network by enabling dropout at test time, one can derive mathematically grounded uncertainty estimates.

We performed an experiment this method to evaluate our CNN's confidence when assessing context-dependent quality in different types of inputs (Fig. 12). We found that our model's predictive uncertainty is lowest for high-frequency images most prone to visible distortion (*e.g.,* Fig. 12b) and very low-frequency images least prone to visible distortion (*e.g.,* Fig. 12a). In images like Fig. 12a where color is uniform or there is not significant detail throughout the patch, compression artifacts are difficult to notice and a scientist might perceive the quality to be good even though the image is significantly compressed. Conversely, in high-detail images such as Fig. 12b, compression artifacts are most noticeable and a scientist might perceive the quality of the image to be poor even though minimal compression was applied to the image. These observations are consistent with the low uncertainty of the model in these image categories. The predictive uncertainty of our model is highest for medium-frequency images (*e.g.,* Fig. 12c) where distortion might be moderate depending on the context of the observation. In Fig. 12c, a large part of the image is sand where a scientist might not notice compression artifacts but the image does contain some fine detail areas where artifacts are visible. Depending on if these areas were the focus of the original image, the scientist reviewing the image may or may not classify the quality of the image as acceptable. In Fig. 12c, our model predicts that a scientist reviewing this image would classify the quality as not accept-

able, but with significant uncertainty. These are cases where scientists are most uncertain when making decisions about the quality of a scientific image. This human uncertainty is consistent with our model uncertainty.

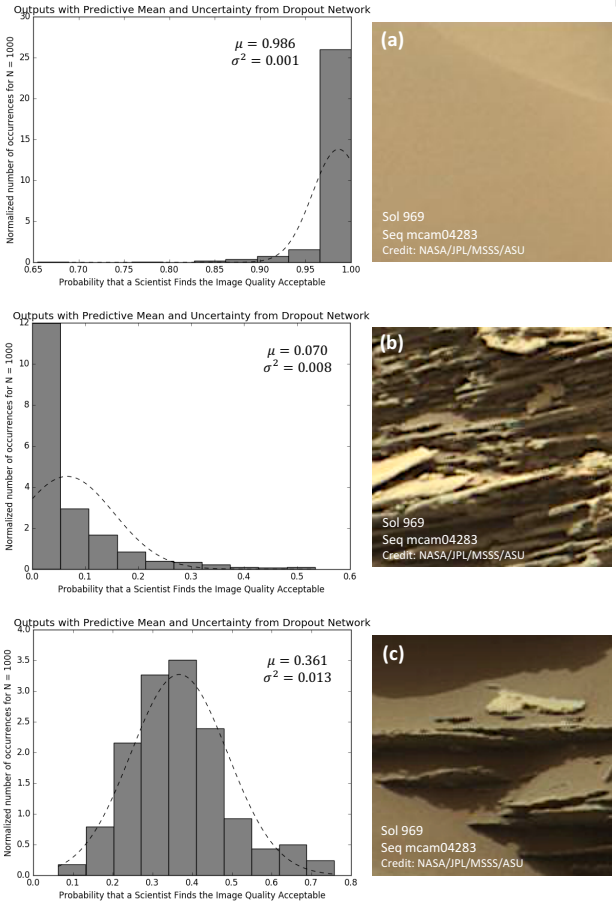


Figure 12: Model outputs and Gaussian fit parameters from model uncertainty evaluation experiment based on Gal and Ghahramani (2016) for three input image examples.

7. Discussion

In general, previous works proposed general solutions to the no-reference image quality assessment problem that are tested on benchmark image quality datasets and do not take into account the user's understanding of the image content. We propose a solution for context-dependent image quality assessment to estimate the perceived quality of JPEG-compressed images that also depends on the scientific context of features in the image. This work also differs from previous work in that we do not perform any pre-processing on images other than slicing full-resolution images into 160×160-pixel image patches for local assessment. Common datasets for developing no-reference image quality assessment models are typically small, artificially distorted, and/or labeled through crowd-sourcing platforms. For our model and application, we need a large dataset with domain-specific labels that cannot be obtained through crowd-sourcing platforms. To solve this problem, we also propose an automatic labeling system based on joint entropy that takes a small number of examples labeled by a domain expert and fits a model that is used to label thousands more examples.

Rather than measuring an objective image quality score, our method estimates the scientific utility as perceived by a scientist using the data. In this work, the level of distortion as well as the scientific context of the image observation determine whether an image should be retransmitted losslessly or not. Our method is designed specifically for the generally well-understood distortions caused by JPEG compression and is trained end-to-end with the Mastcam image dataset. The greatest challenge in developing a context-dependent image quality assessment method is the subjectivity of the science quality interpretations of users. To improve the results presented here, future work could explore incorporating the scientific intent explicitly in the training examples and CNN predictions (rather than implicitly as in this work) or training an ensemble of models that individually model the different scientific interests of users. User-specific models could then be used when planning Mastcam observations to select the appropriate JPEG compression quality for the user that requested

the observation.

8. Conclusions

In this paper, we introduce a new process called *context-dependent* image quality assessment in which the context and intent behind the image observation define the acceptable image quality threshold. We proposed a two-part machine learning solution to estimate the image quality of compressed Mastcam images given the context of the observation as perceived by a scientist. This differs from previous work on image quality analysis because quality is not an objective measure of feature distortion in an image, but rather a context-specific measure of scientific utility that represents the likelihood that artifacts introduced during compression will complicate the scientific analysis of the image.

First, a logistic regression model based on joint entropy between a compressed and uncompressed version of an image was trained using a small set of data to predict the label (accept or retransmit) a scientist might apply to an image if both the compressed and uncompressed image were available. This method enabled us to label a large enough dataset to train a CNN without requiring domain experts to label tens of thousands of images.

Second, we use this labeled data to train a CNN to estimate the scientific utility of a compressed image, or the probability that a scientist using Mastcam data would accept the quality of a compressed image given the observation's context. We demonstrated with a user study that the proposed CNN's predictions correlate with perceptions of context-dependent image quality by scientists. When assisted by our proposed method, we conclude that a Mastcam scientist could spend significantly less time reviewing a subset of images prioritized by our machine learning method than with the existing manual method that requires the investigator to review all of the images.

Acknowledgments

This work was in part supported by NASA funding from the Mars Science Laboratory Mastcam instrument investigation.

References

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28]

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015).

[2] J. F. Bell, A. Godber, S. McNair, M. A. Caplinger, J. N. Maki, M. T. Lemmon, J. Van Beek, M. C. Malin, D. Wellington, K. M. Kinch, M. B. Madsen, C. Hardgrove, M. A. Ravine, E. Jensen, D. Harker, R. B. Anderson, K. E. Herkenhoff, R. V. Morris, E. Cisneros, R. G. Deen, The Mars Science Laboratory *Curiosity* rover Mastcam instruments: Preflight and in-flight calibration, validation, and data archiving, Earth and Space Science 4 (7) (2017) 396–452. doi:10.1002/2016EA000219.

[3] S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, Signal, Image and Video Processing (2017) 1–8doi:10.1007/s11760-017-1166-8.

[4] A. Chetouani, A. Beghdadi, S. Chen, G. Mostafaoui, A Novel Free Reference Image Quality Metric Using Neural Network Approach, in: Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM, 2010.

[5] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning (2016).

[6] D. Ghadiyaram, A. C. Bovik, Blind image quality assessment on real distorted images using deep belief nets, in: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2014, pp. 946–950. doi:10.1109/GlobalSIP.2014.7032260.

[7] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: AISTATS '11: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Vol. 15, 2011, pp. 315–323. arXiv:1502.03167, doi:10.1.1.208.6449.

[8] J. P. Grotzinger, J. Crisp, A. R. Vasavada, R. C. Anderson, C. J. Baker, R. Barry, D. F. Blake, P. Conrad, K. S. Edgett, B. Ferdowski, R. Gellert, J. B. Gilbert, M. Golombek, J. Gómez-Elvira, D. M. Hassler, L. Jandura, M. Litvak, P. Mahaffy, J. Maki, M. Meyer, M. C. Malin, I. Mitrofanov, J. J. Simmonds, D. Vaniman, R. V. Welch, R. C. Wiens, Mars Science Laboratory Mission and Science Investigation, Space Science Reviews 170 (1-4) (2012) 5–56. doi:10.1007/s11214-012-9892-2.

[9] W. Hou, X. Gao, D. Tao, X. Li, Blind Image Quality Assessment via Deep Learning, IEEE Transactions on Neural Networks and Learning Systems 26 (6) (2015) 1275–1286. doi:10.1109/TNNLS.2014.2336852.

[10] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional Neural Networks for No-Reference Image Quality Assessment, IEEE Conference on Computer Vision and Pattern Recognition (2014) 1733–1740doi:10.1109/CVPR.2014.224.

[11] L. Kang, P. Ye, Y. Li, D. Doermann, Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 2791–2795. doi:10.1109/ICIP.2015.7351311.

[12] G. A. Korn, T. M. Korn, Mathematical handbook for scientists and engineers : definitions, theorems, and formulas for reference and review, Dover Publications, 2000.

[13] D. P. Kingma, J. L. Ba, Adam: a Method for Stochastic Optimization, in: International Conference on Learning Representations 2015, 2015, pp. 1–15. arXiv:1412.6980, doi:http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503.

[14] L. Chaofeng, A. C. Bovik, W. Xiaojun, Blind Image Quality Assessment Using a General Regression Neural Network, IEEE Transactions on Neural Networks 22 (5) (2011) 793–799. doi:10.1109/TNN.2011.2120620.

[15] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, Multimodality Image Registration by Maximization of Mutual Information, IEEE Transactions on Medical Imaging 16 (2).

[16] M. C. Malin, M. A. Ravine, M. A. Caplinger, F. Tony Ghaemi, J. A. Schaffner, J. N. Maki, J. F. Bell, J. F. Cameron, W. E. Dietrich, K. S. Edgett, L. J. Edwards, J. B. Garvin, B. Hallet, K. E. Herkenhoff, E. Heydari, L. C. Kah, M. T. Lemmon, M. E. Minitti, T. S. Olson, T. J. Parker, S. K. Rowland, J. Schieber, R. Sletten, R. J. Sullivan, D. Y. Sumner, R. Aileen Yingst, B. M. Duston, S. McNair, E. H. Jensen, The Mars Science Laboratory (MSL) Mast cameras and Descent imager: Investigation and instrument descriptions, Earth and Space Sciencedoi:10.1002/2016EA000252.

[17] A. Mittal, A. K. Moorthy, A. C. Bovik, No-Reference Image Quality Assessment in the Spatial Domain, IEEE Transactions on Image Processing 21 (12) (2012) 4695–4708. doi:10.1109/TIP.2012.2214050.

[18] A. K. Moorthy, A. C. Bovik, Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality, IEEE Transactions on Image Processing 20 (12) (2011) 3350–3364. doi:10.1109/TIP.2011.2147325.

[19] M. A. Saad, A. C. Bovik, C. Charrier, Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain, IEEE Transactions on Image Processing 21 (8) (2012) 3339–3352. doi:10.1109/TIP.2012.2191563.

[20] H. R. Sheikh, Z. Wang, L. Cormack, A. C. Bovik, LIVE Image Quality Assessment Database Release 2.

[21] R. Soundararajan, A. C. Bovik, RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment, IEEE Transactions on Image Processing 21 (2) (2012) 517–526. doi:10.1109/TIP.2011.2166082.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.

[23] H. Tang, N. Joshi, A. Kapoor, Blind Image Quality Assessment Using Semi-supervised Rectifier Networks, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 2877–2884. doi:10.1109/CVPR.2014.368.

[24] Y. Peng, J. Kumar, L. Kang, D. Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1098–1105. doi:10.1109/CVPR.2012.6247789.

[25] Z. Wang, H. Sheikh, A. Bovik, No-reference perceptual quality assessment of JPEG compressed images, in: Proceedings of the International Conference on Image Processing, Vol. 1, IEEE, 2002, pp. I–477–I–480. doi:10.1109/ICIP.2002.1038064.

[26] Z. Wang, E. P. Simoncelli, Reduced-Reference Image Quality Assessment Using A Wavelet-Domain Natural Image Statistic Model, Human Vision and Electronic Imaging 5666 (January 2005) (2005) 149–159.

[27] J. Wu, W. Lin, G. Shi, A. Liu, Reduced-Reference Image Quality Assessment With Visual Information Fidelity, IEEE Transactions on Multimedia 15 (7) (2013) 1700–1705. doi:10.1109/TMM.2013.2266093.

[28] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Advances in Neural Information Processing Systems 27 (NIPS '14), 2014. arXiv:1411.1792.

Highlights

- Convolutional neural network used to predict image quality given scientific context
- Method for automatically labeling images based on joint entropy information loss
- Stochastic pass experiments show model uncertainty reflects human uncertainty
- Automatic image quality analysis can save significant time for instrument P.I.s