Automated Resolution Scoring of Ground-Based LEO Observations Using Convolutional Neural Networks

Jacob Lucas, Trent Kyono, Michael Werth
The Boeing Company
Justin Fletcher
Odyssey Systems Consulting
Ian McQuaid
AFRL

Abstract

The Space-object National Imagery Interpretability Rating Scale (SNIIRS) allows human analysts to provide a quantitative score of image quality based on identification of target features. It is naturally difficult to automate this scoring process, not only because the scale is based on identifiable features but also because the images may be in an almost-resolved image quality regime that is difficult to handle for traditional machine vision techniques. In this paper we explore using a convolutional neural network to automatically produce SNIIRS scores. The neural network is trained with a catalog of analyst-graded images of resolved space objects and then its performance is assessed by comparing the network accuracy to that of a trained analyst.

1. Introduction

This paper explores the viability of augmenting a human-in-the-loop image quality grading task with a Convolutional Neural Network (CNN). The recently illustrated superiority of CNNs at performing a majority of computer vision tasks (described in section 2) implies that they would be well suited to the task of grading image quality. The utility of space object images is constrained by the resolution and quality of the image. The Space-object National Imagery Interpretability Rating Scale (SNIIRS), a variant of the National Imagery Interpretability Scale (NIIRS) [1], allows human analysts to provide a quantitative score of image quality anchored on the size of resolvable target features. The task of scoring currently requires a human analyst due to the almost-resolved regime in which many of the images reside, as well as the score dependence on resolved feature size rather than signal to noise ratio (SNR), point spread function size [2], etc. The nature of the SNIIRS metric leads to high degree of subjectivity, as what constitutes a 'resolved feature' can vary between equally trained and capable analysts.

The SNIIRS rating scale ranges from 0-12, and while scores are typically defined and reported as integer values, nothing precludes finer resolution scoring. This allows us to approach the problem as a regression task. We refer to the regression task as scoring throughout this paper. We evaluate a range of ImageNet pre-trained models and grade by performance in terms of mean absolute error (MAE) for the task of scoring given a single image frame. The task of scoring from a single input frame is assumed to be more challenging than the same task given an image ensemble as input.

We present experiments performed on a substantial set of simulated data using realistic renders of real satellites. The purpose of this study is to evaluate the utility of CNNs at the task of scoring for the purpose of improving score standardization, reducing human-in-the-loop reliance, and reducing time to score. Scoring

in real time would allow for real-time optimization of collection settings during image acquisition, potentially boosting overall image quality.

We provide a discussion of related works in Section 2. Section 3 details a formalization of our problem and approach. Section 4 provides a discussion of our dataset, training architecture, hyperparameters, and our experimental results. We conclude with brief remarks in Section 5.

2. Related Works

Several related works have explored the applications of CNNs and deep learning to astronomy. For noise reduction, [3] recently presented a proof-of-concept neural network for denoising the bispectrum for astronomical image recovery on synthetic data. For classification, [4, 5, 6] investigated the application of object classification using neural networks on photometric light curves and showed promising results. Using Generative Adversarial Networks (GANs), [7] recovered features from artificially degraded images with worse seeing and higher noise than the original with a performance that far exceeded the capabilities of simple deconvolution. Additionally, [8] used a GAN to generate more realistic images of galaxies than existing state of the art. [9] used machine learning to automatically segment and label galaxies in astronomical images. [10] showed promising results using an autoencoder for real-time MFBD of solar images. Additionally there have been probes into image scoring with deep learning; [11] applies a CNN to images and yields a 'human opinion' quality score, and [12] uses a relatively shallow network to give a quality score to distorted images. The success of these approaches motivates our application of similar networks to the scoring of ground based LEO (Low Earth Orbit) observations.

3. Formalization

Let $\mathcal{X}$ and $\mathcal{Y}$ be two spaces, where $\mathcal{X}$ is a set of simulated astronomical images corresponding to a collection, and $\mathcal{Y}$ is the collections *score*, i.e., $\mathcal{Y} = \{\mathbb{R}\}$ (regression). We refer to each collection $x_i \in \mathcal{X}$ as a pass containing n sequential images. Given a pass $x_i \in \mathcal{X}$, our primary design goal is to train a network $f : \mathcal{X} \rightarrow \mathcal{Y}$, which takes as input a collection $x_i \in \mathcal{X}$ and provides a *scoring* prediction, $y_i \in \mathcal{Y}$. To do this we train a single image classifier $g : \mathcal{X}_j \rightarrow \mathcal{Y}_j$ that takes as input a single image, $x_{ij} \in \mathcal{X}_j$, for the j^{th} image belonging to a pass x_i and makes a prediction $y_{ij} \in \mathcal{Y}_j$ for image x_{ij}. We then repeat this action for every image in pass x_i.

4. Experiments

This section briefly covers our datasets used in this work, our experimental settings (training architecture and regimes), and our experimental results.

4.1. Datasets

The SNIIRS score of a space object is a measure of the smallest resolvable feature. It is a log_2 based score, with a larger value indicating that smaller features are resolvable. An analyst is required to determine what features are resolvable in a given image, and to determine the size of the smallest resolvable feature. Possible scores range from 0 (no resolvable features) to 12 (features smaller than 5mm can be resolved). In practice many of the higher scores are not currently attainable for ground based observations. The useable range of this metric for ground based systems can be viewed in terms of r_0 and object range. A reasonable span of r_0 values might be from 5cm (poor seeing) to 100cm (adaptive optics corrected). At a typical LEO range of 600km, these r_0 values equate to resolutions of approximately 10m (SNIIRS 3) and .5m (SNIIRS

7) respectively in the I-band. For this study a significant database of analyst-scored ground-based LEO observations was made available, however this data contained a lack of diversity that even with class weighting was difficult if not impossible to overcome, with several classes having no representation in the data at all. The solution posed was to construct a simulated basis set for initial training, and then for future work fold in actual sensor data.

To create the simulated dataset a metric was constructed using a validated atmospheric simulation code [13] to apply a broad range of atmospheric turbulence characterized by Fried Parameter r_0 to a scaled 3-bar target render with cascading target sizes (Fig. 1). By setting the simulated range and instantaneous field of view (IFOV), we designated the simulated physical size of the 3-bar targets. Measuring the smallest resolvable 3-bar target for a simulated r_0 and then fitting in the style of a General Image Quality Equation (GIQE) [14] produced a mapping from $\Delta\ r_0$ to Δ SNIIRS. Simulating a reasonable set of r_0 values allowed us to degrade an image by multiple SNIIRS with a resolution of approximately 1/4 of a SNIIRS. In the simulation used r_0 was a dependent variable and could not be directly set, only measured. This resulted in range of SNIIRS scores with less recognizable but still entirely valid values, as can be viewed in Figure 2. We constrained our scores to a range from 3 to 7, corresponding to r_0 of 5cm and 100cm as described in the previous paragraph.

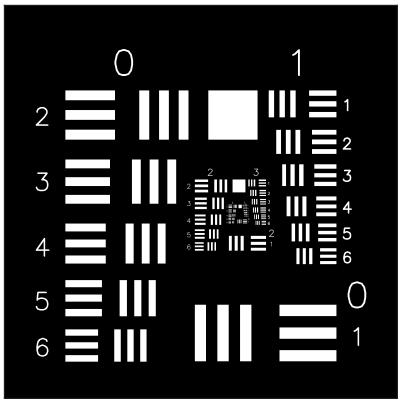

Figure 1: 3-bar target used to calibrate the simulation SNIIRS scores.

An image set containing 5000 satellite renders with 38 discrete satellites in multiple poses and configurations was used as the basis for the simulated set. The SNIIRS score of each render was established by a trained analyst. This is referred to as the absolute score. The absolute score is the SNIIRS of the target as viewed from a diffraction-limited optical system with a circular aperture, i.e. the upper limit for the simulated SNIIRS is capped at the absolute score on the basis of resolvable features discenable in the diffraction limited image. The basis set contains a range of initial SNIIRS values, which after degradation provides a diverse data set of greater than 90,000 scenarios each containing 125 simulated images. From here on this simulated data will be referred to as SILO (Simulated Images of LEO Objects).

A subset of SILO was used for network training. This subset was evenly sampled across SNIIRS 3-7 as shown in Fig. 2. To prevent contamination of the validation data, the subset was segregated by target render. Of the 38 discrete targets, the training set contained 30, and the remaining 8 were reserved for validation and testing (discrete sets). This ensured that images in the validation data were dissimilar to the training images.

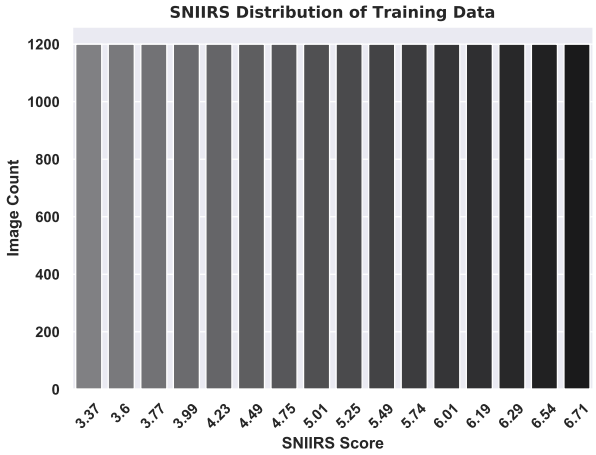

Figure 2: Histogram illustrating the distribution of SNIIRS scores in the simulated training data.

80% of data was used for training, 20% for validation and testing.

4.2. Performance

In this work, we investigated the following networks: Densenet 121/169/201 [15], Inception Resnet V2 [16], Inception V3 [17], NasNet Large and Mobile [18], ResNet 50 [19], VGG 16/19 [20], and Xception [21]. For each of the pre-trained models we replaced the top dense layers with a following sequence of top layers (non-convolutional layers, * following layer indicates *ReLu* activation followed by a dropout layer with rate of 0.2): a global average pooling layer, a 1024 neuron dense layer*, a 256 neuron dense layer*, and finally a 1 neuron dense layer with *linear* activation. Each dense layer was initialized with Glorot normal distribution [22].

For input image augmentation, we applied random augmentation to each training image with the following specification: rotation within ± 30 degrees, horizontal flips, vertical flips, and zoom within $\pm 20\%$. Image augmentation was not used during testing inference. We used an iterative deepening training regime, where we first train the top dense layers by freezing all lower layers (CNN layers) for 5 epochs. We then freeze the top dense layers and "unfreeze" the base model and train for 5 epochs. This is followed by another round of 5 epochs with base model frozen and the top dense layers trainable. Lastly we make the entire model trainable for 10 epochs. We use the *Adam* optimizer with a learning rate of 10^{-3} and 10^{-4} for the first and last two iterations, respectively. We used a Mean Absolute Error (MAE) loss function, and saved models at the point with which the validation loss did not improve.

The predictions on our test set are presented in Table 1. We show our results both with and without Contrast Limited Adaptive Histogram Equalization (CLAHE) augmentation. CLAHE augmentation is a method shown in [23] to improve classification performance, particularly when issuing predictions on grayscale images on pre-trained ImageNet models that are optimized for 8-bit red-green-blue (RGB) inputs. Formally, the CLAHE grid size and nominal clip limit was augmented according to the following:

$$a \in \mathcal{U}(-log_2(k), log_2(k)) \mid g(k) = k + a, \tag{1}$$

where k is the nominal grid size or the nominal clip limit [23]. In Table 1, we see that while CLAHE augmentation improved our predictive performance as a whole, the best performing network in terms of MAE

Table 1: Performance of various ImageNet models on our dataset with or without CLAHE augmentation. Bold denotes model best performance. VGG16 and VGG19 failed to converge with the training scheme as attempted.

Model	**Without CLAHE**	**With CLAHE**
	MAE	MAE
Densenet 121	0.276	0.187
Densenet 169	0.405	0.180
Densenet 201	0.255	**0.126**
Inception Resnet V2	0.140	0.283
Inception V3	0.253	0.252
NasNet Large	1.130	1.527
NasNet Mobile	1.069	0.458
ResNet 50	0.368	0.251
VGG 16	0.938	0.938
VGG 19	0.939	0.938
Xception	**0.124**	0.140

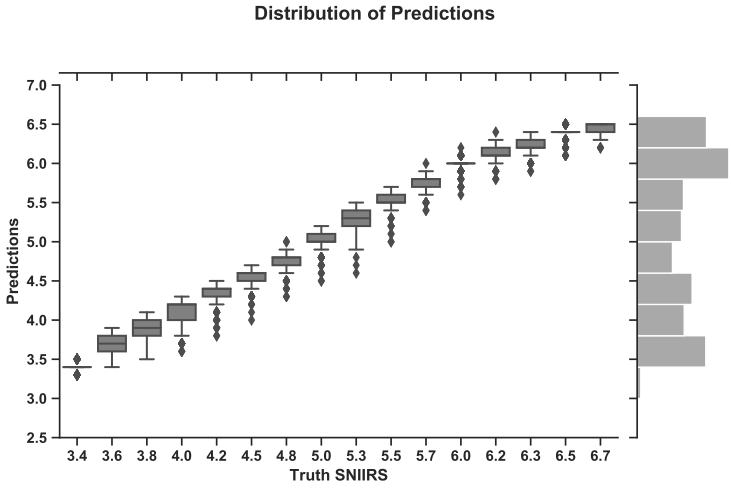

Figure 3: Distribution of 2560 test set predictions from the peak scoring network evenly sampled over the full range of SNIIRS scores. Histogram at right displays density of predictions.

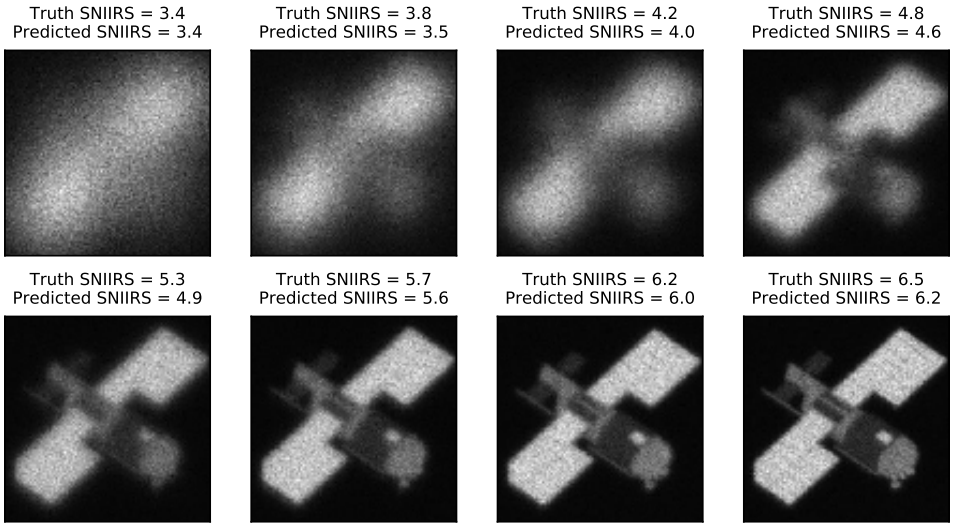

Figure 4: Example satellite render with range of applied degradations. SNIIRS scores are typically reported as integers. The values here have been rounded to one decimal to illustrate network performance. The predictions shown were obtained with the peak scoring network from Table 1.

was Xception trained without CLAHE, this network is used for all predictions shown herein. Adjustments to the training regime described earlier in this section could change this result. Figure 3 looks at the distribution of network predictions vs truth scores for all images from the test set; i.e. objects never before seen by the network. Figure 4 is a mosaic of a particular render at varying SNIIRS scores with comparisons between prediction and truth.

4.3. Reproducibility

For reproduction, all neural networks were trained using Python 3 and Keras in conjunction with TensorFlow [24, 25]. Operating system and hardware specifications include RedHat Linux 7 on an NVidia DGX Workstation with four Tesla V100 GPUs with 32 GB of memory on each card. Because the goal of this work was a feasibility investigation, we did not tune or search for optimal hyperparameters.

5. Conclusion

We have shown that CNNs can perform admirably at the heretofor human only task of scoring LEO image quality. Particularly, given a single frame from our dataset we showed that the Xception network performed the task optimally, and that CLAHE augmentation generally improved performance. For future work, we plan to test network performance on real data to determine transferability, followed by optimization of real data performance by training with an amalgam of real and simulated images. It may also be beneficial to explore a customized network, as well as perform an exhaustive search for optimal hyperparameters. Additionally, we would like to apply active learning to improve and sharpen our overall *scoring* performance.

References

[1] John M. Irvine. National imagery interpretability rating scales (niirs): overview and methodology, 1997.

[2] M. Werth, J. Bos, B. Calef, S. Williams, D. Thompson, and S. Williams. A new performance metric for hybrid adaptive optics systems. In *2014 IEEE Aerospace Conference*, pages 1–11, March 2014.

[3] Jacob Lucas, Brandoch Calef, and Trent Kyono. Recovering astronomical images with deep neural network supported bispectrum processing. In *Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*. 2018.

[4] B. Jia, K. D. Pham, E. Blasch, Z. Wang, D. Shen, and G. Chen. Space object classification using deep neural networks. In *2018 IEEE Aerospace Conference*, pages 1–8, March 2018.

[5] Richard Linares. Space object classification using deep convolutional neural networks. 07 2016.

[6] Robert Furfaro, Richard Linares, and Vishnu Reddy. Space objects classification via light-curve measurements: Deep convolutional neural networks and model-based transfer learning. 09 2018.

[7] Kevin Schawinski, Ce Zhang, Hantian Zhang, Lucas Fowler, and Gokula Krishnan Santhanam. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. *Monthly Notices of the Royal Astronomical Society: Letters*, 467(1):L110–L114, 01 2017.

[8] Levi Fussell and Ben Moews. Forging new worlds: high-resolution synthetic galaxies with chained generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*, 485(3):3203–3214, 03 2019.

[9] Alex Hocking, James E. Geach, Yi Sun, and Neil Davey. An automatic taxonomy of galaxy morphology using unsupervised machine learning. *Monthly Notices of the Royal Astronomical Society*, 473(1):1108–1129, 09 2017.

[10] A. Asensio Ramos, J. de la Cruz Rodriguez, and A Pastor Yabar. Real-time multiframe blind deconvolution of solar images. 06 2018.

[11] Hossein Talebi Esfandarani and Peyman Milanfar. NIMA: neural image assessment. *CoRR*, abs/1709.05424, 2017.

[12] Le Kang, Peng Ye, Yi Ci Li, and David S. Doermann. Convolutional neural networks for no-reference image quality assessment. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[13] Thomas C. Farrell Jacob Lucas V. S. Rao Gudimetla, Richard B. Holmes. Phase screen simulations of laser propagation through non-kolmogorov atmospheric turbulence. In *Proc. SPIE, Atmospheric Propagation VIII*, volume 8038, 2011.

[14] Jon C. Leachtenauer, William Malila, John Irvine, Linda Colburn, and Nanette Salvaggio. General image-quality equation: Giqe. *Appl. Opt.*, 36(32):8322–8328, Nov 1997.

[15] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[16] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[18] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. *CoRR*, abs/1707.07012, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[21] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[22] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[23] Trent Kyono, Fiona J. Gilbert, and Mihaela van der Schaar. MAMMO: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. *CoRR*, abs/1811.02661, 2018.

[24] François Chollet et al. Keras, 2015.

[25] Martín Abadi et.al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.