Metrics for Evaluating Galaxy Image Generators

Bailer Hackstein*

Department of Data Science, University of Applied Sciences Northwestern Switzerland, Windisch, Switzerland

Introduction

Verifying that the generated distribution resembles the target distribution while the individual generated sample is indistinguishable from the original data is a major issue with deep generative models. In particular, for use in astrophysics, we need to make sure that the generated samples include all object types with the right frequency and diversity and that the generated data matches our prior knowledge. Currently, we lack objective methods for systematically evaluating these quality aspects, where human inspection reaches its limits due to the need for in-depth data analysis. In this work, appropriate metrics for the quality of galaxy image generators are identified. A small number of conditional image generators trained on galaxy images with visual morphology feature classification labels are compared for this purpose. A new set of cluster-based metrics for comparing the generated distribution to the target distribution is our main contribution. Additionally, the Wasserstein distance and a number of other common image generator metrics are applied to galaxy morphology proxies [1].

Automated mode collapse identification is made possible by the newly introduced cluster-based metrics, which are excellent proxies for the generated distribution's quality. In addition, the generated distribution can be qualitatively interpreted using the cluster metrics. The morphological statistics-based metrics are a useful tool for determining the physical soundness of generated samples. Last but not least, although it is difficult to interpret, we discover that the kernel inception distance utilized with an ImageNet-pre-trained InceptionV3 model accurately reflects the overall quality of galaxy image generators [2].

Description

Upcoming telescope-based astronomical surveys like Euclid, as well as LSST will supply billions of galaxy images and a wealth of data. The formation and evolution of galaxies, the cosmic distribution of dark matter, and the Universe's expansion history are just a few of the well-studied astrophysical and cosmological issues these address. However, astrophysicists cannot individually investigate the vast number of images because they are far too numerous. Instead, in order to inform and constrain physical models, galaxy properties must be quickly and methodically extracted from their images. The astrophysical community already uses a number of computational tools that have been developed and even though these make it possible to extract properties in a systematic way, they use too much computing power to produce large collections of high-quality mock images in a reasonable amount of time. As a result, it is necessary to use faster methods to replace these tools, such as deep neural networks or machine learning techniques [3].

Before applying machine learning techniques or automated analysis tools to actual data, it is absolutely necessary to test their accuracy on unobserved data sets. Testing a full inference pipeline that combines several physical and deep learning models is a much more elusive endeavor, in contrast to, for instance, testing a classifier model trained on supervised data, which can be easily tested by splitting the training data into training and test sets. To find out if such pipelines are prone to distinguishing between competing physical models, they ought to be

put through their paces in a variety of possible scenarios. In order to accomplish this, a collection of synthetic galaxy images must be produced. Semi-analytical models can be used to generate galaxies from scratch, which enables control over the distribution of the types of generated galaxies. However, these models cannot generate the entire universe's galaxy diversity because they must make simplifying assumptions, such as the galaxy's perfect rotational symmetry [4].

Full-fledged physical numerical simulations, such as cosmological simulations, can also be used to generate the galaxies associated with the large-scale structure by allowing for fine resolution in high density peaks. Because they allow for accurate predictions of competing models of cosmology and galaxy formation as well as realistic propagation effects that affect observations, these simulations produce the most coherent datasets. Unfortunately, they cannot generate large datasets for a sufficient number of competing physical models, which are required for the evaluation of extensive inference pipelines, as they are computationally expensive. Utilizing deep generative models like variational autoencoders (VAE) or generative adversarial networks (GAN) is a promising strategy [5].

The evaluation of generative models presents a significant obstacle in determining the reason for the disparity in statistical characteristics between the generated and target samples. We investigate a comprehensive set of evaluation metrics for galaxy image generators in this work. While some of these metrics are physically motivated and tailored to galaxy images, others are frequently utilized in machine learning. We distinguish helpful measurements for evaluating various parts of the nature of generative models. Physical properties, distribution characteristics, and image quality per sample are all included. We can also use these metrics to evaluate particular generated datasets, like balanced sets or samples that represent opposing physical models. In particular, we present a brand-new collection of cluster-based metrics that qualitatively evaluate a generated set's feature distribution. In order to direct the improvement of generative models, these make it possible to identify mode collapse and problematic object types. It is important to note that their use transcends galaxy images [2].

We employ RGB images of galaxies from the Sloan Digital Sky Survey for our research, which the Galaxy Zoo data challenge provided. A citizen science approach has been used to collect the label information. We need a quick and dependable tool to automatically classify the visual morphology of generated galaxy images for conditional training on these labels. This will enable the definition of an additional loss term, which will enhance the generator's training. Dieleman et al., earlier works examined the use of various techniques, primarily convolutional neural networks (CNNs), to build a deep neural network that can classify visual galaxy morphology features with high accuracy. A BigGAN-based model and two conditional deep generative models are used in conjunction with this morphological image classifier and a straightforward conditional VAE. The InfoSCC-GAN, develope, in which a different classifier and encoder are used to classify visual galaxy morphology features instead. Finally, we make use of a collapsed generator with the typical mode. We demonstrate that the evaluation metrics presented in this work enable us to determine which of the competing models generates the most high-quality samples [5].

Using generative models in physics presents a significant challenge in ensuring that the generated images accurately reflect the variety and quality of actual data. For the case of galaxies in particular, we need to produce physically sound images of every kind of galaxies that have been observed throughout the universe, including their various shapes, morphologies, and image quality. In addition, we require techniques for ensuring that a set of generated galaxies accurately reflects the physical scenario encoded in the input variables. We investigate a number of metrics for evaluating the performance of galaxy image generators in this work [4].

RGB images with labels for visual morphology features from a citizen science project make up our dataset. The Galaxy Zoo datasets are the only ones that, as far as we are aware, contain labels for a number of visual galaxy morphological

*Address for Correspondence: Bailer Hackstein, Department of Data Science, University of Applied Sciences Northwestern Switzerland, Windisch, Switzerland, E-mail: hackstein_b@gmail.com

features like elasticity, spiral arms, bars, and bulges. It is impossible to locate the original survey data because the images have been anonymized from the SDSS data. As a result, we must utilize RGB images rather than raw survey data. Notwithstanding, according to the computational perspective, this is a benefit, as required computational assets can be significantly diminished by the utilization of pre-handled information. In addition, the RGB images contain all of the necessary information for morphological classification. With the exception of the morphological proxies, the presented evaluation metrics do not depend on the actual format of the dataset. As a result, using raw survey data rather than RGB images will not benefit this work in any way. However, all of the methods and findings presented in this work can be applied to raw survey data, providing a proof of concept [3].

To prevent differences in sampling, we use identical splitting into training, validation, and test sets for the training of all neural networks. We can thus isolate and concentrate on the alterations in results brought about by various model architectures. Although a cross-validation strategy would lessen the splitting-induced bias, it would also take more time to compute. Nevertheless, the method we select permits systematic comparison of competing models. The evaluation metrics presented in this paper can be used to compare their quality. The primary focus of this work is the investigation of various evaluation metrics and their capacity to evaluate various galaxy image generator quality aspects. To evaluate these metrics, it's best to compare models with obvious benefits and disadvantages. As a result, our work does not make use of fully optimized models. Instead, an extensive ablation study we are currently conducting will include the optimization of the presented models [1].

Conclusion

We investigate a number of evaluation metrics that have the potential to measure various quality aspects in order to find suitable evaluation metrics for evaluating the quality of galaxy image generators. A small number of conditional generative models, some deliberately of lower quality than others, are used to investigate these metrics. We discovered evaluation metrics that are useful proxies for the quality of individual images and the target distribution's similarity.

Acknowledgement

Conflict of Interest

None.

References

1. Bretonnière, H., A. Boucaud, F. Lanusse and E. Jullo, et al. "Euclid preparation-XIII. Forecasts for galaxy morphology with the Euclid Survey using deep generative models." *Astron Astrophys* 657 (2022): A90.

2. Castelvecchi, Davide. "Astronomers explore uses for AI-generated images." *Nat* 542 (2017): 16-17.

3. Conselice, Christopher J. "The relationship between stellar light distributions of galaxies and their formation histories." *Astrophys J Supplement Series* 147 (2003): 1.

4. Dieleman, Sander, Kyle W. Willett and Joni Dambre. "Rotation-invariant convolutional neural networks for galaxy morphology prediction." *Mon Notices Royal Astron Soc* 450 (2015): 1441-1459.

5. Boquien, M., D. Burgarella, Y. Roehlly and V. Buat, et al. "CIGALE: a python code investigating GALaxy emission." *Astron Astrophys* 622 (2019): A103.