Latent space representations of cosmological fields

Sambatra Andrianomena
South African Radio Astronomy Observatory
Department of Physics & Astronomy,
University of the Western Cape
andrianomena@gmail.com

Sultan Hassan [*]
New York University
Flatiron Institute
University of the Western Cape
sultan.hassan@nyu.edu

Abstract

We investigate the possibility of learning the representations of cosmological multifield dataset from the CAMELS project. We train a very deep variational encoder on images which comprise three channels, namely gas density (Mgas), neutral hydrogen density (HI), and magnetic field amplitudes (B). The clustering of the images in feature space with respect to some cosmological/astrophysical parameters (e.g. Ω_m) suggests that the generative model has learned latent space representations of the high dimensional inputs. We assess the quality of the latent codes by conducting a linear test on the extracted features, and find that a single dense layer is capable of recovering some of the parameters to a promising level of accuracy, especially the matter density whose prediction corresponds to a coefficient of determination $R^2 = 0.93$. Furthermore, results show that the generative model is able to produce images that exhibit statistical properties which are consistent with those of the training data, down to scales of $k \sim 4h/\mathrm{Mpc}$.

1 Introduction

The data deluge which will be coming from current and different surveys at different wavelengths, e.g. DESI [Aghamousa et al., 2016], SKA [Dewdney et al., 2009, Weltman et al., 2020] and HIRAX [Newburgh et al., 2016], will improve our understanding of the large-scale structure. In order to uncover patterns and anomalies in the data from those surveys, advanced techniques will be required. Deep networks that are considered for parameter inference at the field level can help extract the salient features from high dimensional data (e.g. 2D images). Although cosmological simulations are useful for training those convolutional neural network (CNN) based methods to recover the underlying cosmology and astrophysics [Gupta et al., 2018, Ribli et al., 2019, Fluri et al., 2019, Villaescusa-Navarro et al., 2021a, Andrianomena and Hassan, 2023], extracting the relevant features from noisy data in real world scenario still proves to be challenging for those networks. One way to tackle that issue is to train the networks with noisy mock data [e.g. Hassan et al., 2020], making them more robust when predicting inputs in the presence noise. However, systematics from real surveys can be hard to model, which still poses a problem. Another approach, which is less sensitive to complex noise, is to learn the representations of the high dimensional data. In general, representation learning which consists of extracting the salient features from the inputs does not rely on the labels of the latter and is suitable for discovering patterns and anomalies. Self-supervised learning [Chen et al., 2020, Grill et al., 2020, He et al., 2020, Chen and He, 2021] has been proven to be able to successfully extract useful representations of images (e.g. CIFAR10). Our main objective in this work is to learn the latent codes of the multifield images which are composed of three 2D maps (channels) of different physical fields. This is crucial in preparation for analyzing and understanding the unlabelled data from different cosmological surveys. In contrast with self-supervised learning based approaches, our method – a variant of variational autoencoder (VAE) – is simultaneously

---

[*]NASA Hubble Fellow

trained to extract important information from the input, and generate *new* images that have both the quality and statistical properties that are comparable to the those of the training data. Previous works also used VAE models to analyze cosmological data [e.g. Tröster et al., 2019, Yi et al., 2020, Thorne et al., 2021, Horowitz et al., 2022].

2 Method

2.1 Data

We consider the public CAMELS Multifields Dataset (CMD) [Villaescusa-Navarro et al., 2022] which was obtained from running state-of-the art hydrodynamic simulation IllustrisTNG from the CAMELS Project [Villaescusa-Navarro et al., 2021b] . In CMD, there are 13 different fields representing a region of $25 \times 25 \times 5 \ (h^{-1}\mathrm{Mpc})^3$ at $z = 0$. In this proof of concept, each instance in our training/testing dataset comprises three channels corresponding to gas density (Mgas), neutral hydrogen (HI) and magnetic fields magnitude (B). The three different fields (or channels) in each example describe the same physical region with the same underlying cosmological Ω_m and σ_8 and astrophysical parameters (A_SN1, A_SN2, A_AGN1, A_AGN2) characterizing the supernova and active galactic nuclei feedbacks. The training, validation and test sets contain 13,000, 1,000 and 1,000 maps of 256×256 pixels^2 respectively.

2.2 Algorithm

In our analyses, we consider the variational autoencoder prescribed in Child [2020]. VAE [Kingma and Welling, 2013] comprises an encoder $q_\phi(z|x)$, a decoder $p_\theta(x|z)$ and a prior $p_\theta(z)$. The training of the two networks ϕ and θ consists of maximizing the evidence lower bound (ELBO) [Kingma and Welling, 2013]

$$\mathbb{ELBO} = E_{z \sim q_\phi(z|x)} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x)||p_\theta(z)), \tag{1}$$

where the first term is the reconstruction loss, whereas the second term is the Kullback-Leibler divergence, measuring the dissimilarity between the approximate posterior $q_\phi(z|x)$ and the prior $p_\theta(z)$. Child [2020] introduced a hierarchical VAE, which is composed of many stochastic layers of latent variables, such that both the approximate posterior and the prior, which generate the latent variables, are more expressive. The latent variables which are conditionally dependent on each other (like in autoregressive models) have different resolutions according to

$$p_\theta(z) = p_\theta(z_0) \prod_{k=1}^{N} p_\theta(z_k|z_{k-1}),$$

$$q_\phi(z|x) = q_\phi(z_0|x) \prod_{k=1}^{N} q_\phi(z_k|z_{k-1}, x), \tag{2}$$

where N is the number of layers, and $q_\phi(\cdot)$ and $p_\theta(\cdot)$ are diagonal Gaussian distributions. It is worth noting that, unless otherwise stated, the learned representations (or latent codes) in our analyses denote the latent variables with the lowest resolutions (vector of length 256). The encoder comprises many stages of stacked residual blocks and the downsampling after the last residual block in one stage is achieved with an average pooling. A residual block has 4 convolutional layers. A top-down block uses one residual block for the prior, one residual block for the posterior, one convolutional layer for the latent variable and a last residual block at its ouput. The decoder has the same number of stages as the encoder, and each of them is composed of stacked top-down blocks. The upsampling after the last top-down block at a given stage is done by using nearest neighor (see the schematic diagram in Figure 3 of Child [2020]). During training, the conditionals in equation 2 are achieved by mixing (via concatenation) the output of the last residual block at a given stage of the encoder with the inputs of all top-down blocks at the corresponding stage of the decoder. We consider 6 stages with (3, 3, 2, 2, 2, 2) number of residual blocks for the encoder, and (3, 3, 2, 2, 2, 2) number of top-down blocks for the decoder. We train the model for 300 epochs using RMSprop optimizer with learning rate = 0.00002, momentum = 0.9 and weight decay = 10^{-4}, in batches of 4 instances on a NVIDIA GeForce GTX 1080 Ti. Each epoch takes about 26 minutes.
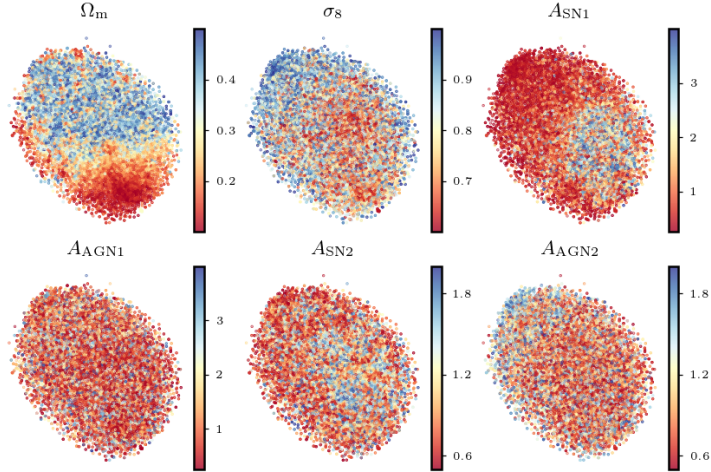
Figure 1: Visualization of the learned representations using t-SNE. The color coding in each panel denotes the parameter values.
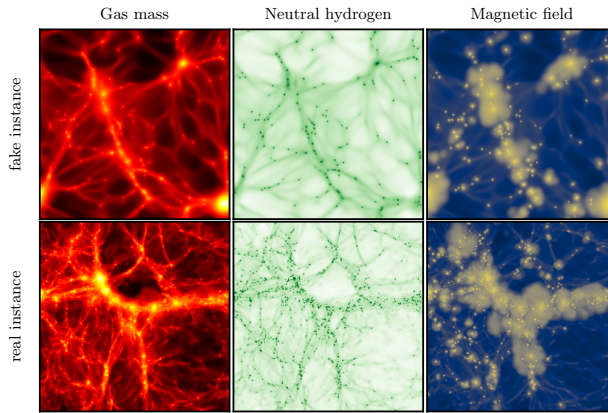

Figure 2: Comparing field images of a new instance generated by the trained decoder on the top row with those of a real instance at the bottom row from the test set. Each field of both real and fake instances is shown in the same column.

3   Results

To investigate the learned latent codes during training, the whole dataset is fed into the encoder in order to extract the features which consist of vectors of length 256. For visualization, the features are further projected into a two dimensional subspace using a dimensionality reduction method, t-distributed stochastic neighbor embedding (t-SNE). Figure 1 shows the clustering of the data points in 2D space. Each point in each panel is color-coded according to the value of the underlying parameter, and each panel presents the clustering with respect to one parameter, e.g. Ω_{m}. Overall, the clustering of the latent representations is visible with respect to the cosmological parameters (Ω_{m}, σ_8) and the supernova feedback parameters (A_{SN1}, A_{SN2}), although the trend related to the AGN feedback parameters is very weak or non-existent. In other words, feature vectors associated with roughly the same value of any of those four parameters tend to be clumped together in the embedding space. This is indicative of the ability of the generative model to learn the latent codes of the inputs.

For the linear test, the new dataset composed of the features extracted (vector of length 256) from the entire dataset is similarly split into training, validation and testing with 13,000, 1,000 and 1,000 examples respectively. A single dense layer without any activation is trained and tested on the learned representations for 200 epochs, using Adam optimizer with a learning rate of 0.0025. We
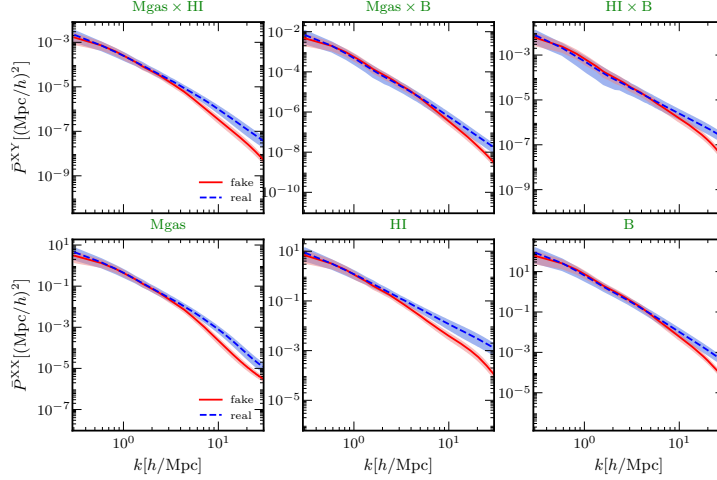
3

Figure 3: Top row: average cross-power spectrum between the fields for both the test images (dashed blue) and generated images (solid red) as a function of wavenumber k. Bottom row: average auto-power spectrum of each field in the test images (dashed blue) and in the fake images (solid red). In all cases, the shaded area denotes the standard deviation of a power spectrum, blue and red for the real and fake respectively.

use coefficient of determination R^2 to assess the performance of the single layer network, and obtain $R^2 = 0.93, 0.64, 0.72, -0.01, 0.85, 0.20$ for Ω_m, σ_8, A_SN1, A_AGN1, A_SN2 and A_AGN2 respectively. Results suggest that the latent space representations of the images still carries relevant information for the feed-forward neural network to recover some parameters to a relatively good accuracy ($R^2 > 0.7$). The recovery of the matter density Ω_m corresponds to the highest accuracy $R^2 = 0.93$, which is consistent with the most apparent clustering of data points of similar Ω_m values. The challenge related to the predictions of the AGN feedback parameters was also highlighted in Villaescusa-Navarro et al. [2021a].

From sampling in the latent space, the model is capable of generating new examples which are similar to the training dataset using the decoder. Field images of a real instance from the test set are qualitatively compared with those of a new instance generated by the decoder in Figure 2. Visually, the quality of the generated images, exhibiting the topological features like voids, filament, halos and bubbles, is comparable with that of the data. However the fake images tend to be slightly blurry, which points toward the fact that capturing the small scale features is challenging to our model. To investigate the statistical properties of the generated instances, we compute: the average ($\bar{P}^\mathrm{XX}(k)$) and standard deviation ($\sigma_{P^\mathrm{XX}}(k)$) of auto-power spectrum of each channel of all images in the test set, and both average ($\bar{P}^\mathrm{XY}(k)$) and standard deviation ($\sigma_{P^\mathrm{XY}}(k)$) of the cross-power spectrum between channels of all the test instances. We plot $\bar{P}^\mathrm{XX}(k)$ for each field at the bottom row of Figure 3. The real and fake power spectra are in good agreement down to $k \sim 4h/\mathrm{Mpc}$. This comes to corroborate the fact that the generated images are a bit blurred, causing their power spectra to be lower than those of the real images on smaller scales. The consistency between the average cross-power spectra between channels of the real and generated instances up to $k \sim 5h/\mathrm{Mpc}$ (Figure 3 top row) indicates that the channels of the generated samples are consistent with one another, i.e. the three field images describe the same physical region. The clustering properties of the fake maps demonstrate that the generative model is capable of producing maps that mimic the training data.

4   Conclusion

We have demonstrated in this work that meaningful latent space representations of CAMELS multi-field dataset can be learned using Very Deep Variational AutoEncoder (VDVAE), which is a generative model. The latent codes, which still carry useful information, can be used for some downstream task such as recovering the cosmological and some astrophysical parameters of the maps to a relatively good accuracy. The new maps generated by the model are in good agreement with the data up to a

level of high frequency, as evidenced by the consistency between the cross/auto-power spectra of fake and real images for a wavenumber $k \leq 5h/\mathrm{Mpc}$. This limitation of the model can be addressed by further fine-tuning the decay of the KL-divergence part of the loss function, or increasing the model capacity or the combination of the two. For future work, a normalizing flow based model can be trained on the latent codes. This will provide the likelihood of the inputs, which is very important for parameter inference or out-of-distribution task. In order to investigate the robustness of the learned representations, the trained model can be tested on different datasets, such as SIMBA.

Acknowledgments and Disclosure of Funding

References

Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, Christophe Balland, Otger Ballester, et al. The desi experiment part i: science, targeting, and survey design. *arXiv:1611.00036*, 2016.

Peter E Dewdney, Peter J Hall, Richard T Schilizzi, and T Joseph LW Lazio. The square kilometre array. *Proceedings of the IEEE*, 97(8):1482–1496, 2009.

Amanda Weltman, P Bull, S Camera, K Kelley, H Padmanabhan, J Pritchard, A Raccanelli, S Riemer-Sørensen, L Shao, S Andrianomena, et al. Fundamental physics with the square kilometre array. *Publications of the Astronomical Society of Australia*, 37:e002, 2020.

LB Newburgh, K Bandura, MA Bucher, T-C Chang, HC Chiang, JF Cliche, R Davé, M Dobbs, C Clarkson, KM Ganga, et al. Hirax: a probe of dark energy and radio transients. In *Ground-based and Airborne Telescopes VI*, volume 9906, pages 2039–2049. SPIE, 2016.

Arushi Gupta, José Manuel Zorrilla Matilla, Daniel Hsu, and Zoltán Haiman. Non-gaussian information from weak lensing data via deep learning. *Physical Review D*, 97(10):103515, 2018.

Dezső Ribli, Bálint Ármin Pataki, and István Csabai. An improved cosmological parameter inference scheme motivated by deep learning. *Nature Astronomy*, 3(1):93–98, 2019.

Janis Fluri, Tomasz Kacprzak, Aurelien Lucchi, Alexandre Refregier, Adam Amara, Thomas Hofmann, and Aurel Schneider. Cosmological constraints with deep learning from kids-450 weak lensing maps. *Physical Review D*, 100(6):063514, 2019.

Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N Spergel, Yin Li, Benjamin Wandelt, Andrina Nicola, Leander Thiele, Sultan Hassan, Jose Manuel Zorrilla Matilla, et al. Multifield cosmology with artificial intelligence. 2021a.

Sambatra Andrianomena and Sultan Hassan. Predictive uncertainty on astrophysics recovery from multifield cosmology. *Journal of Cosmology and Astroparticle Physics*, 2023(06):051, 2023.

Sultan Hassan, Sambatra Andrianomena, and Caitlin Doughty. Constraining the astrophysics and cosmology from 21 cm tomography using deep learning with the ska. *Monthly Notices of the Royal Astronomical Society*, 494(4):5761–5774, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

Tilman Tröster, Cameron Ferguson, Joachim Harnois-Déraps, and Ian G McCarthy. Painting with baryons: augmenting n-body simulations with gas using deep generative models. *Monthly Notices of the Royal Astronomical Society: Letters*, 487(1):L24–L29, 2019.

Kai Yi, Yi Guo, Yanan Fan, Jan Hamann, and Yu Guang Wang. Cosmo vae: Variational autoencoder for cmb image inpainting. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

Ben Thorne, Lloyd Knox, and Karthik Prabhu. A generative model of galactic dust emission using variational autoencoders. *Monthly Notices of the Royal Astronomical Society*, 504(2):2603–2613, 2021.

Benjamin Horowitz, Max Dornfest, Zarija Lukić, and Peter Harrington. Hyphy: Deep generative conditional posterior mapping of hydrodynamical physics. *The Astrophysical Journal*, 941(1):42, 2022.

Francisco Villaescusa-Navarro, Shy Genel, Daniel Angles-Alcazar, Leander Thiele, Romeel Dave, Desika Narayanan, Andrina Nicola, Yin Li, Pablo Villanueva-Domingo, Benjamin Wandelt, et al. The camels multifield data set: Learning the universe's fundamental parameters with artificial intelligence. *The Astrophysical Journal Supplement Series*, 259(2):61, 2022.

Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N Spergel, Rachel S Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, et al. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, 2021b.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv:2011.10650*, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.