HISTOGRAM POOLING OPERATORS: AN INTERPRETABLE ALTERNATIVE FOR DEEP SETS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper we describe the use of a differentiable histogram as a pooling operator in a Deep Set. For some set summarization problems, we argue this is the more interpretable choice compared to traditional aggregations, since it has functional resemblance to manual set summarization techniques. By staying closer to standard techniques, one can make more explicit connections with the learned functional form used by the model. We motivate and test this proposal with a large-scale summarization problem for cosmological simulations: one wishes to predict global properties of the universe via a set of observed structures distributed throughout space. We demonstrate comparable performance to a sum-pool and mean-pool operation over a range of hyperparameters for this problem. In doing so, we also increase the accuracy of traditional forecasting methods from 20% error using our dataset down to 13%, strengthening the argument for using such methods in Cosmology. We conclude by using our operator to symbolically discover an optimal cosmological feature for cosmic voids (which was not possible with traditional pooling operators) and discuss potential physical connections.

1 INTRODUCTION

Data-rich scientific fields such as astrophysics have increasingly found success in applying machine learning methods to different parts of the research process (Cranmer et al., 2020; Ntampaka et al., 2021). The most common strategy is to exploit the powerful high-dimensional pattern-finding ability of deep learning, and to model relationships in large observational datasets which would be too difficult to discover by hand. However, the natural sciences are fundamentally about the understanding of these models rather than simply producing accurate predictions. We want to bridge this divide between the powerful predictive ability of deep learning and the interpretation of learned models. Achieving this will give deep learning the ability to contribute to our fundamental understanding of the natural sciences.

In this paper, we focus on the following problem. We are interested in summarizing a set—be it a dataset of celestial structures, a point cloud of particles, or any other data structure which can be represented as a set of vectors. We would like to take a set of vectors, and reduce the set into a single vector. The standard way of approaching this problem is with a technique known as "Deep Sets" (Zaheer et al., 2017) A Deep Set can be thought of as a specialization of a Graph Neural Network (see figure 4 of Battaglia et al., 2018), leaving out the edges which create explicit interactions (Battaglia et al., 2016; Santoro et al., 2017). The simplest functional form for a DeepSet is:

$$\boldsymbol{y} = f(\ \rho\left(\{g(\boldsymbol{x}_i)\}\right)\ ), \tag{1}$$

for a set of vectors $\{\boldsymbol{x}_i\}_{i=1:N}$, permutation-invariant pooling operator ρ , learned functions f and g, and output vector $\boldsymbol{y}$. These algorithms typically use traditional pooling operators, such as element-wise summation, averaging, or max (Goodfellow et al., 2016). This turns out to be an incredibly flexible algorithm, capable of accurately modelling quite an impressive range of datasets (e.g., Komiske et al., 2019; Shlomi et al., 2020; Oladosu et al., 2020). For certain problems, one can even interpret the pooled vectors as physical variables, such as forces (Cranmer et al., 2019; 2020), if the pooling operation has a strong connection to the physical problem (e.g., summed force vectors).

However, for some systems, such as the dataset we are interested in, none of these pooling operations match the canonical form. Here, we argue that a much more interpretable pooling operation for summarizing a set is the one-dimensional *histogram* applied independently to each latent feature. An informal argument motivating this is to note that when someone approaches a new dataset in machine learning, the first thing they might do to understand the data is to view the histogram of different features. In a more formal and practical sense, the histogram is a very common operation in science (and much of statistics in general) for describing a set of objects. In cosmology, the central tool for reducing observational data to a statistic is a histogram of distances between objects—the "correlation function"—which is used to make predictions about the initial state of the universe.

So, though slightly more complex, one can argue that this operator may well be much more interpretable, given its common occurrence in so many existing hand-engineered and physical models, allowing one to compare the learned latent features to theory. In this work, we demonstrate this new pooling operation for Deep Sets—histogram pooling—and show that it is much more interpretable than the sum/mean/max-pool for our dataset, while not sacrificing performance or changing the number of parameters. We first describe our dataset and particular Cosmology problem of interest. We then give a mathematical statement for our pooling operator, and compare it with traditional pooling operators on our dataset.

2 DATASET

The distribution of galaxies in the Universe follows a complex, web-like pattern called the cosmic web. Different components of the cosmic web such as galaxies or cosmic voids—vast under-dense regions—can be used to extract information about our Universe. Such information includes "Cosmological parameters," which describe the various physical properties of the universe such as matter content.

Cosmic voids are becoming increasingly important to use in analyses (see Pisani et al., 2019, for a review). While current observations already allow for the extraction of cosmological constraints from voids (e.g., Hamaus et al., 2020), theoretical models for these objects have not yet matured (Pisani et al., 2015; Verza et al., 2019; Contarini et al., 2020; Kreisch et al., 2019).

In this work we utilize simulations of cosmic voids in the universe to: (1) illustrate the interpretability of our new histogram pooling operation and (2) also to provide a new theoretical framework for void analyses.

We use 2000 simulations from the QUIJOTE simulations set (Villaescusa-Navarro et al., 2020), a suite of 44,100 full N-body cosmology simulations. Each simulation is ran with a different set of cosmological parameters, allowing us to compute the inverse problem with machine learning. We identify cosmic voids and their features in the simulations by running the VIDE[1] software (Sutter et al., 2015). For further details on the simulation data and void finding, see Appendix A. In Figure 1 we show an example of the spatial distribution of the underlying matter from one of our simulations with the white ellipses representing just a few of the underdense regions identified as cosmic voids. We split our 2,000 void catalogs into three sets: training (70%), validation (10%) and testing (20%).

3 MODEL

Standard operators used in DeepSets and Graph Neural Networks include sum-pool, average-pool, and max-pool, which perform an element-wise aggregation over vectors. Here, we describe a a pooling operator that, instead of an aggregation like $\mathbb{R}^n \to \mathbb{R}$, performs the aggregation $\mathbb{R}^n \to \mathbb{R}^m$ for m bins. Recent work has also considered pools which similarly map to a non-scalar number of features, such as sort-based pools (Zhang et al., 2019). Our aggregation, after applied to every element, can be followed by a flattening so that each bin of each latent feature becomes an input feature to the multilayer perceptron.

A version of a sum-pool Deep Set can be described by the following operations, to map from a set $\{\boldsymbol{x}_i\}$ to a summary vector $\boldsymbol{y}$:

$$\boldsymbol{y} = f(\sum_i \boldsymbol{z}_i) \quad \text{for} \quad \boldsymbol{z}_i = g(\boldsymbol{x}_i) \tag{2}$$

[1] `https://bitbucket.org/cosmicvoids/vide_public`

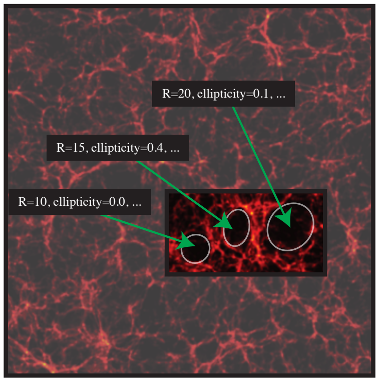
R=20, ellipticity=0.1, ...
R=15, ellipticity=0.4, ...
R=10, ellipticity=0.0, ...

Figure 1: Field in the Quijote simulations (Villaescusa-Navarro et al., 2020) with a few voids circled and labeled.

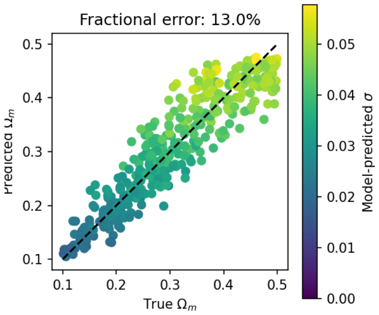

Figure 2: Value and error estimates for Ω_{m} using the best trained histogram model.

for learned functions f and g, which in our case, are multi-layer perceptrons. Our proposed pooling operation changes this to:

$$\boldsymbol{y} = f(\text{flatten}(\boldsymbol{w})) \quad \text{where} \tag{3}$$

$$w_{jk} = \sum_i e^{-(a_k - z_{ij})^2/2\sigma^2}, \quad \text{for} \quad \boldsymbol{z}_i = g(\boldsymbol{x}_i) \tag{4}$$

where z_{ij} is the j-th latent feature of element i, a_k is a hyperparameter giving a pre-defined bin position for bin k, σ is a hyperparameter controlling the histogram's smoothness, w_{jk} is the histogram value for feature j and bin k, and $\boldsymbol{w}$ is a matrix with its j-th row and k-th column as w_{jk}. Again, f and g are learned functions. Since we smoothly weight the contributions to the histogram, this operation preserves gradient information backpropagated to the parameters of g.

In this case of the histogram pool, f would receive an input vector which is m times larger than in the sum-pool case, essentially allowing a fewer number of latent features to communicate more information. To compare these models at a given number of latent features, we increase the width of the sum-pool model's multi-layer perceptron hidden layers until both models have the same number of parameters.

4 Experiments

We demonstrate our model by learning to predict the value of "Ω_{m}," fraction of the total energy density that is matter at the present time in the universe, from cosmological simulations. Increasing the accuracy at which we can estimate this parameter and others from real-world observations is one of the central directions in modern cosmology research.

We predict the mean and variance of a Gaussian distribution over Ω_{m} and train with log-likelihood. Since we work in a finite box in parameter space, we use the two-sided truncated Gaussian loss to remove model bias to this parameter box.

We train models over a range of number of latent features (number of outputs of the function g), and compare between pooling operations (see Appendix B). The sum and average pool have their hidden size increased so they have the same number of parameters as the histogram pool.

Experiments are summarized in Figure 3. As can be seen, the performance is similar among all three types of pooling, as they are at almost within about 1σ of each other at each choice of number of latent features. From this, we argue that there may be cases such as our example dataset where this type of pooling does not hurt performance, while greatly improve interpretability.

Following this, we attempt to recover the learned latent feature in the histogram-pooling case for a single latent feature. This is directly comparable to the typical model used for void cosmology. Our model achieves about 13% error on predictions for Ω_{m}, quite an improvement over a classic

Fisher forecast that achieves 20% error (Kreisch et al., 2021). We compare our predictions to truth in Figure 2.

5 Explicit Interpretation

The standard approach for estimating cosmological parameters using voids is to first compute the following function:

$$g(R) = n(> R),$$

where $n(> R)$ is the number density of voids which have their radius larger than some value. The resultant cumulative histogram passed through f is traditionally being a cosmological maximum likelihood analysis.

With our histogram-pool Deep Set, we could learn this relationship easily even for only a single latent feature, since the function g could select R from the input vectors $\boldsymbol{x}_i$, and the function f would then emulate the regular cosmology "analysis pipeline."

Our best model yields the following interpretable equation for g in Equation 4:

$$z_{i1} = -\alpha R_i + \beta \delta_{\mathrm{c}i} - \gamma R_i \epsilon_i + C, \tag{5}$$

$$\text{where} \quad \alpha = 0.17,\ \beta = 0.57,\ \gamma = 0.026,\ \text{and}\ C = 0.16, \tag{6}$$

and where R_i, $\delta_{\mathrm{c}i}$, and ϵ_i represent the void radius, density contrast, and ellipticity, respectively, for void i, after normalizing each feature over all simulations to have unit variance and zero mean. Density contrast indicates the ratio between the field density outside the void and inside the void, and can be thought of as the void's depth. The equation provides an optimal feature to constrain cosmological parameters. The first term captures the trend of large voids having less constraining power on Ω_{m}. The second term highlights the importance of a void's density contrast, while the third indicates a weak, though non-trivial, dependence on ellipticity. The constant suggests additional astrophysical parameters not yet included may contribute to enhanced constraints on Ω_{m}. In the literature, there has not yet been a clear connection allowing to directly constrain Ω_{m} with the density contrast. This work indicates that there is a strong, straightforward link between the density contrast and Ω_{m} to be used for void modeling, also opening up the exploration of the impact of different astrophysical parameters, like galaxy bias.

6 Discussion

Our work demonstrates an interpretable pooling operation for sets using a one-dimensional histogram. We achieve similar performance to standard pooling operations while allowing for the learned model to be interpretable in the language of manual set summarization techniques. Our proposed pooling operation is capable of capturing higher order moments in the condensed data distribution with a single feature (hence, easier to interpret), whereas standard operations like sum-pooling would need to learn as many latent features as there are moments to achieve this.

We have illustrated this interpretability by applying our model to a large set of cosmic void simulations. For our problem of interest in Cosmology, our histogram-pooling results in the discover of a a new relationship between the amount of matter in the universe and a cosmic void's size, shape, and depth, which potentially holds large significance for cosmology research. This work suggests the possibility of symbolically interpreting Deep Set models with the right inductive bias, in this case through an introduction of a novel pooling operation. In the future we plan to test this pooling operation on a more general set of problems.

References

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction Networks for Learning about Objects, Relations and Physics. *arXiv e-prints*, art. arXiv:1612.00222, December 2016.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv e-prints*, art. arXiv:1806.01261, June 2018.

Sofia Contarini, Federico Marulli, Lauro Moscardini, Alfonso Veropalumbo, Carlo Giocoli, and Marco Baldi. Cosmic voids in modified gravity models with massive neutrinos. *arXiv e-prints*, art. arXiv:2009.03309, September 2020.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1912789117. URL https://www.pnas.org/content/117/48/30055.

Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering Symbolic Models from Deep Learning with Inductive Biases. *arXiv e-prints*, art. arXiv:2006.11287, June 2020.

Miles D. Cranmer, Rui Xu, Peter Battaglia, and Shirley Ho. Learning Symbolic Physics with Graph Networks. *arXiv e-prints*, art. arXiv:1909.05862, September 2019.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Nico Hamaus, Alice Pisani, Jin-Ah Choi, Guilhem Lavaux, Benjamin D. Wandelt, and Jochen Weller. Precision cosmology with voids in the final BOSS data. , 2020(12):023, December 2020. doi: 10.1088/1475-7516/2020/12/023.

Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy flow networks: deep sets for particle jets. *Journal of High Energy Physics*, 2019(1):121, January 2019. doi: 10.1007/JHEP01(2019)121.

Christina D. Kreisch, Alice Pisani, Carmelita Carbone, Jia Liu, Adam J. Hawken, Elena Massara, David N. Spergel, and Benjamin D. Wandelt. Massive neutrinos leave fingerprints on cosmic voids. , 488(3):4413–4426, September 2019. doi: 10.1093/mnras/stz1944.

Christina D. Kreisch, Alice Pisani, Francisco Villaescusa-Navarro, and David N. Spergel. Quijote void catalogs. In preparation, 2021.

Michelle Ntampaka, Camille Avestruz, Steven Boada, Joao Caldeira, Jessi Cisewski-Kehe, Rosanne Di Stefano, Cora Dvorkin, August E. Evrard, Arya Farahi, Doug Finkbeiner, Shy Genel, Alyssa Goodman, Andy Goulding, Shirley Ho, Arthur Kosowsky, Paul La Plante, Francois Lanusse, Michelle Lochner, Rachel Mandelbaum, Daisuke Nagai, Jeffrey A. Newman, Brian Nord, J. E. G. Peek, Austin Peel, Barnabas Poczos, Markus Michael Rau, Aneta Siemiginowska, Danica J. Sutherland, Hy Trac, and Benjamin Wandelt. The role of machine learning in the next decade of cosmology, 2021.

Ademola Oladosu, Tony Xu, Philip Ekfeldt, Brian A. Kelly, Miles Cranmer, Shirley Ho, Adrian M. Price-Whelan, and Gabriella Contardo. Meta-Learning for Anomaly Classification with Set Equivariant Networks: Application in the Milky Way. *arXiv e-prints*, art. arXiv:2007.04459, July 2020.

Alice Pisani, P. M. Sutter, Nico Hamaus, Esfandiar Alizadeh, Rahul Biswas, Benjamin D. Wandelt, and Christopher M. Hirata. Counting voids to probe dark energy. , 92(8):083531, October 2015. doi: 10.1103/PhysRevD.92.083531.

Alice Pisani, Elena Massara, David N. Spergel, David Alonso, Tessa Baker, Yan-Chuan Cai, Marius Cautun, Christopher Davies, Vasiliy Demchenko, Olivier Doré, Andy Goulding, Mélanie Habouzit, Nico Hamaus, Adam Hawken, Christopher M. Hirata, Shirley Ho, Bhuvnesh Jain, Christina D. Kreisch, Federico Marulli, Nelson Padilla, Giorgia Pollina, Martin Sahlén, Ravi K. Sheth, Rachel Somerville, Istvan Szapudi, Rien van de Weygaert, Francisco Villaescusa-Navarro, Benjamin D. Wandelt, and Yun Wang. Cosmic voids: a novel probe to shed light on our Universe. , 51(3):40, May 2019.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv e-prints*, art. arXiv:1706.01427, June 2017.

Jonathan Shlomi, Peter Battaglia, and Jean-Roch Vlimant. Graph Neural Networks in Particle Physics. *arXiv e-prints*, art. arXiv:2007.13681, July 2020.

P. M. Sutter, G. Lavaux, N. Hamaus, A. Pisani, B. D. Wandelt, M. Warren, F. Villaescusa-Navarro, P. Zivick, Q. Mao, and B. B. Thompson. VIDE: The Void IDentification and Examination toolkit. *Astronomy and Computing*, 9:1–9, March 2015. doi: 10.1016/j.ascom.2014.10.002.

Giovanni Verza, Alice Pisani, Carmelita Carbone, Nico Hamaus, and Luigi Guzzo. The void size function in dynamical dark energy cosmologies. , 2019(12):040, December 2019. doi: 10.1088/1475-7516/2019/12/040.

Francisco Villaescusa-Navarro, ChangHoon Hahn, Elena Massara, Arka Banerjee, Ana Maria Delgado, Doogesh Kodi Ramanah, Tom Charnock, Elena Giusarma, Yin Li, Erwan Allys, Antoine Brochard, Cora Uhlemann, Chi-Ting Chiang, Siyu He, Alice Pisani, Andrej Obuljen, Yu Feng, Emanuele Castorina, Gabriella Contardo, Christina D. Kreisch, Andrina Nicola, Justin Alsing, Roman Scoccimarro, Licia Verde, Matteo Viel, Shirley Ho, Stephane Mallat, Benjamin Wandelt, and David N. Spergel. The Quijote Simulations. , 250(1):2, September 2020. doi: 10.3847/1538-4365/ab9d82.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep Sets. *arXiv e-prints*, art. arXiv:1703.06114, March 2017.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. FSPool: Learning Set Representations with Featurewise Sort Pooling. *arXiv e-prints*, art. arXiv:1906.02795, June 2019.

A APPENDIX: SIMULATIONS & VOID CATALOGS

The QUIJOTE simulations (Villaescusa-Navarro et al., 2020) are a suite of 44,100 full N-body simulations spanning thousands of different cosmologies in the hyperplane $\{\Omega_\mathrm{m}, \Omega_\mathrm{b}, h, n_s, \sigma_8, M_\nu, w\}$, where these parameters respectively represent the amount of matter in the universe, the amount of baryons in the universe, the Universe's expansion speed, the tilt of what we call the primordial power spectrum, how clumpy the universe is, the mass of neutrino particles, and if there is any deviation in the dark energy equation of state. For this work, we make use of the high-resolution latin-hypercube simulations at $z = 0$ (present day). This set is comprised of 2,000 simulations, each following the evolution of 1024^3 dark matter particles from $z = 127$ (early in the universe) down to $z = 0$ (present day) in a periodic comoving volume of $(1000\,h^{-1}\mathrm{Mpc})^3$. The value of the cosmological parameter in that set is arranged into a latin-hypercube with dimensions $\Omega_\mathrm{m} \in [0.1 - 0.5]$, $\Omega_\mathrm{b} \in [0.03 - 0.07]$, $h \in [0.5 - 0.9]$, $n_s \in [0.8 - 1.2]$, $\sigma_8 \in [0.6 - 1.0]$. Dark matter halos, dense spheres of dark matter that form the building blocks of galaxies, in these simulations have been identified using the Friends-of-Friends (FoF) algorithm.

From the halo catalogs we identify cosmic voids by running the VIDE software (Sutter et al., 2015). VIDE produces a catalog of voids, for each simulation, containing properties of the cosmic voids such as positions, volumes, density contrasts, and ellipticities.

B APPENDIX: HYPERPARAMETER TUNING

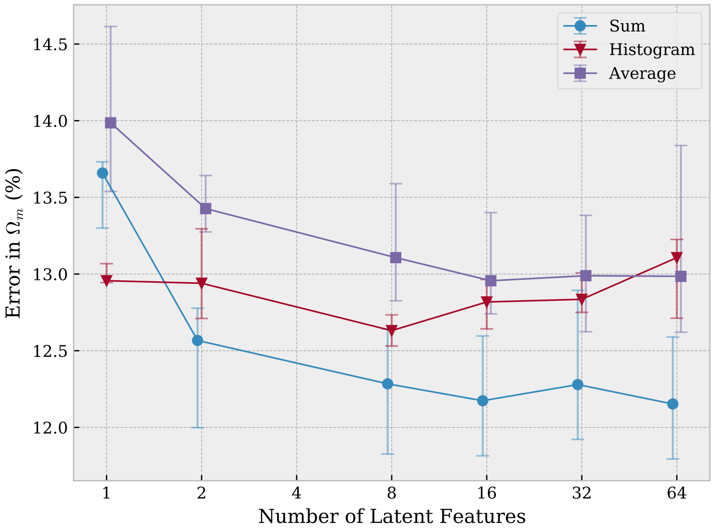

Figure 3: Summary of our experiments comparing sum-pool, average-pool, and histogram-pool on our dataset. The points show the median error over 10 random seeds trained with a fixed number of steps. Error bars show the $[25, 75]$ confidence interval. Slight x-axis perturbations are introduced to the points for ease of visualization.

Our networks for f and g are both 2-layer multi-layer perceptrons, with ReLU activations. We set the histogram positions, a_k, equally spaced in a grid between -1 and $+1$. We fix the number of histogram bins at 64, with the smoothing set to $\sigma = 0.1$. We train our models for 100,000 steps of batch size 4. We train with cosine annealed learning rate with a peak learning rate of 3×10^{-4}, and with gradient clipping at 1.0 applied to the L2-norm of the gradients. For the histogram-pool models, we use a hidden size of 300. For the sum-pool and average-pool, we use a larger hidden size that results in an equal number of parameters as the histogram-pool model, since the number of inputs to g in the histogram-pool is 64 times larger.

We perform a hyperparameter search in terms of the number of latent features over all models, running 10 seeds in each case and computing the median and confidence interval. This is shown in fig. 3.