Resource Efficient and Generalizable Representation Learning of High-Dimensional Weather and Climate Data

Juan Nathaniel*
Columbia University
jn2808@columbia.edu

Marcus Freitag
IBM Research
mfreitag@us.ibm.com

Patrick Curran
Environment and Climate Change Canada
patrick.curran@ec.gc.ca

Isabel Ruddick
Environment and Climate Change Canada
isabel.ruddick@ec.gc.ca

Johannes Schmude
IBM Research
johannes.schmude@ibm.com

Abstract

We study self-supervised representation learning on high-dimensional data under resource constraints. Our work is motivated by applications of vision transformers to weather and climate data. Such data frequently comes in the form of tensors that are both higher dimensional and of larger size than the RGB imagery one encounters in many computer vision experiments. This raises scaling issues and brings up the need to leverage available compute resources efficiently. Motivated by results on masked autoencoders, we show that it is possible to use sampling of subtensors as the sole augmentation strategy for contrastive learning with a sampling ratio of $\sim$1%. This is to be compared to typical masking ratios of 75% or 90% for image and video data respectively. In an ablation study, we explore extreme sampling ratios and find comparable skill for ratios as low as $\sim$0.0625%. Pursuing efficiencies, we are finally investigating whether it is possible to generate robust embeddings on dimension values which were not present at training time. We answer this question to the positive by using learnable position encoders which have continuous dependence on dimension values.

1 Introduction

Following numerous breakthroughs in deep learning, there has been a growing trend to adapt techniques used in computer vision [1, 2] to other domains, including in weather and climate applications [3, 4, 5, 6]. The type of data used, however, is complex due to its high-dimensional nature, especially compared to standard RGB images. For instance, the ERA5 reanalysis [7], spans the globe with extensive number parameters and atmospheric levels, resulting in GB-scale tensors at each time step. This research delves into resource-efficient self-supervised representation learning using ERA5 data, employing invariance-based methods like SimCLR [8, 9] and MoCo [10]. In particular, we use sampling as the sole data augmentation technique and impose resource constraints of 24-hour train time on a single A100 GPU with 40GB memory. Across multiple downstream

*Work done while an intern at IBM Research.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

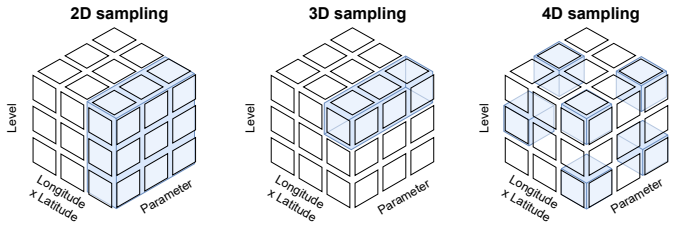

Figure 1: (*left*) Traditional second-order sampling generates a sequence of subtensors by tiling along the two horizontal dimensions. (*center* and *right*) We sample subtensors that extend along all three spatial or even all four dimensions; which we refer to as third- and fourth-order sampling respectively.

regression, classification, and clustering tasks, the learned embeddings generated from our strategy are more robust than previous methods. We also introduce a position encoder that allows our model to generalize to unseen dimension values. In total, our work shows that it is possible to learn skillful representations in an extremely efficient way, which in turn reduces the climate impact of training models for tasks such as weather forecasting [11, 12], extreme event detection [13], and climate model parameterization [14].

2 Our Proposal

2.1 Higher-Order Sampling

While our method generalizes to any weather or climate dataset, we will for concreteness consider ERA5 and consider a single timestamp of a given set of ERA5 parameters on a given set of verticle pressure levels as a sample X^t. The typical approach for invariance-based (contrastive) learning with a vision transformer (ViT) would be to apply some randomized transformations to X^t and then tokenize the sample by cutting it into patches along the horizontal dimensions. This means that the entire data cube is held in GPU memory. Our approach instead patches along either all three spatial dimensions or along all four dimensions. See Figure 1. Then, our only randomized transformation is to sample from this set of patches. I.e. we perform contrastive learning by sampling two sets of tokens from these patches without any further augmentation. Note that we only hold the sampled tokens in GPU memory. Decreasing the sampling ratios reduces GPU memory consumption and makes the training task more difficult. To avoid confusion, let us point out that we run inference without sampling on the complete sample X^t.

2.2 Generalizable Positional Encoder

The second part of our method concerns position encodings. After tokenization, it is best practice to add an additional signal to each token that informs the model about the position of said token in relation to the complete sample. Such encodings can be hardcoded or learned.

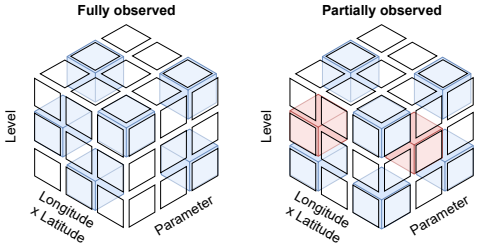

Figure 2: (*left*) Position encodings tend to assume the same dimension values to be present during training and inference. (*right*) By using a position encoder $f_\theta(l)$, we are able to train on a subset of values (blue) yet run inference on the entire set (red and blue) by capturing the non-linear dependencies on coordinates beyond interpolation.

We note that for weather and climate data, many dimensions such as latitude, longitude, the vertical level and time are continuous. Thus, we can either define or learn a position encoding as a function f that depends on these continuous dimension values. Concretely, we will train a model with a learned position encoding $f(\text{level})$. This allows us to train a model on a fixed set of pressure levels yet run inference on another set, without resorting to interpolation. Together with our first proposal, one can train on e.g. 10 pressure levels of ERA5 data yet deploy on all 37.

3 Methodology

3.1 Dataset and Downstream Tasks

We evaluate our proposed methods on the ERA5 reanalysis dataset. For concreteness we restrict to pressure level data at levels 50, 100, 200, 300, 400, 500, 600, 700, 800, 850, 900 and $1000hPa$. At each level, we consider specific humidity (kg/kg), temperature (K), geopotential (gpm), vertical velocity (m/s) as well as the easterly and northerly components of the wind (m/s). We use the years 2000 to 2014 as training and 2015 to 2021 as validation data.

Given learned representations, we perform 1-day forecasting of key atmospheric variables (T850, T1000, Z300, Z500, Z700), classification, and clustering of wildfire hotspots [15, 16] and tropical cyclones [17] using linear probing method. We use RMSE as metric in our regression/forecasting task, and % accuracy for both classification and clustering (i.e., top-1 nearest neighbor in latent space also containing wildfire/cyclone events) tasks.

3.2 Self-Supervised Learning Setup

We end-to-end train ViTs with 12 attention blocks and a single-layer linear projection head. We use both the SimCLR and MoCo frameworks to ensure that our results do not depend on the choice of framework. We specify 768 and 128 as the embedding sizes of the encoder backbone and projection head, respectively. The models are optimized by AdamW [18], and the cosine learning rate policy is used with an initial 5-epoch warm-up. For framework, we use identical specifications as described in [8] and [10] respectively. The second-order sampling baseline is the only approach using view augmentations as specified in the original work [19]: cutout, Sobel filtering, and Gaussian blurs.

4 Results and Discussion

4.1 Higher-Order Sampling is an Efficient and Scalable Augmentation Strategy

Tables 1-2 show the impact of the different sampling strategies on downstream performance. First, we notice that a higher-order sampling tends to generate better downstream performances than the traditional combination of second-order sampling with view augmentation. Our fourth-order sampling produces the most skillful representations (lowest validation RMSE) across all 1-day ahead forecasts of T850, T1000, Z300, Z500, and Z700, and the classification/clustering of wildfire.

		Regression				
Methods	Sampling Ratio	T850	T1000	Z300	Z500	Z700
2D	50%	4.29 / 4.37	4.70 / 4.77	1300 / 1322	830 / 843	486 / 494
3D	$\sim$5%	2.28 / 2.28	2.42 / 2.41	641 / 641	430 / 429	270 / 269
4D (ours)	$\sim$1%	**2.25 / 2.26**	**2.38 / 2.39**	**634 / 637**	**423 / 426**	**267 / 268**

Table 1: Impact of different sampling strategies in downstream regression/1-day forecasting task. The validation RMSE for SimCLR/MoCo are shown.

Given this finding, the immediate and obvious question is how far one can push the sampling approach. To answer this question we reduce the sequence length – and thus the sampling ratio – by factors of four and sixteen. As illustrated in Figure 3, even at the extreme sampling ratio of 0.0625% embeddings are still skillful and in several cases even best. For instance, a $16\times$ reduction in sampling ratio yields embedding that increases wildfire classification accuracy from 84% to 87%, and demonstrate comparable performance across other tasks.

Methods	Sampling Ratio	Classification		Clustering	
		Cyclones	Wildfire	Cyclones	Wildfire
2D	50%	**84.07** / 79.13	82.36 / 80.02	**48.82** / 39.33	62.99 / 61.48
3D	∼5%	80.96 / **82.95**	81.90 / 81.03	47.34 / **46.65**	59.46 / 60.24
4D (ours)	∼1%	83.87 / 78.72	**84.53 / 87.11**	48.52 / 40.46	**63.43 / 65.09**

Table 2: Impact of different sampling strategies in downstream classification and clustering tasks. The validation accuracies in percent for SimCLR/MoCo are shown.

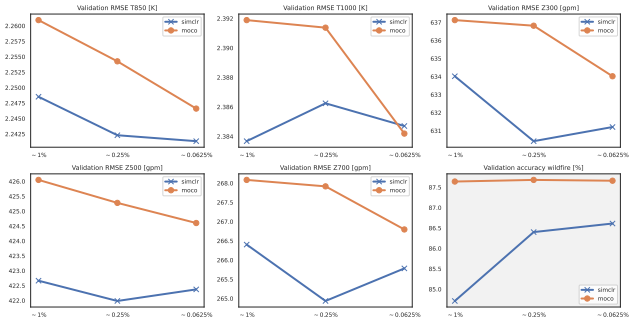

Figure 3: Benchmarks for fourth-order sampling across different sampling ratios. The panel with grey background shows validation accuracy (higher is better); the others RMSE (lower is better).

4.2 Generalizable Positional Encoder

To estimate performance when training only on a subset of dimension values yet running inference on the full set or the unobserved values, we train another set of self-supervised models subject to identical resource constraints. Here, the following pressure levels are not observed at training time: 300, 500 and 700 hPa. In order to evaluate the utility of our proposed position encoder, we compare the skills of embeddings in the fully-observed (baseline) and partially-observed cases (ours). We perform evaluation on similar regression, classification, and clustering downstream tasks.

Training data	Regression		Classification		Clustering	
	Z300	Z500	Cyclones	Wildfire	Cyclones	Wildfire
Full	634 / 637	423 / 426	83.9 / 78.7	84.5 / 87.1	48.5 / 40.5	63.4 / 65.1
Partial	629 / 636	421 / 425	83.8 / 79.4	80.6 / 88.0	44.6 / 38.3	63.4 / 60.2

Table 3: Downstream skill for fully- and partially-observed pressure levels. In the latter case the embedding is inferred using our learned yet continuous position encoder f_θ. The validation RMSE (forecast) and accuracy (classification and clustering) for SimCLR/MoCo are shown.

Our generalization approach in the partially-observed case has comparable downstream performances when benchmarked against the fully-observed baseline case (Table 3). This result is promising since, for one, we are able to end-to-end train a self-supervised model with previously unseen dimension value. This opens up the possibilities to not just (1) train and infer on different data resolution *separately*, but also to (2) train on both multi-resolution datasets *simultaneously*.

5 Conclusion

We propose methods to scale and generalize self-supervised learning in high-dimensional setting by introducing higher-order sampling and learnable, generalizable positional encoder. Our fourth-order sampling effectively reduces sampling ratios to ∼1%, while still producing skillful embeddings that outperform the traditional approach across many downstream tasks. In addition, a further reduction in sampling ratio to ∼0.0625% does not compromise downstream skills. Finally, we demonstrate the utility of our proposed learnable positional encoder by showing similar performance on previously

unobserved levels across all downstream tasks. Overall, both of our proposed methods allow for efficient and scalable self-supervised learning on higher-dimensional data.

References

[1] C. Feichtenhofer, Y. Li, K. He, *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in neural information processing systems*, vol. 35, pp. 35946–35958, 2022.

[2] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

[3] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, *et al.*, "Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators," *arXiv preprint arXiv:2202.11214*, 2022.

[4] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, A. Pritzel, S. Ravuri, T. Ewalds, F. Alet, Z. Eaton-Rosen, *et al.*, "Graphcast: Learning skillful medium-range global weather forecasting," *arXiv preprint arXiv:2212.12794*, 2022.

[5] S. K. Mukkavilli, D. S. Civitarese, J. Schmude, J. Jakubik, A. Jones, N. Nguyen, C. Phillips, S. Roy, S. Singh, C. Watson, *et al.*, "Ai foundation models for weather and climate: Applications, design, and implementation," *arXiv preprint arXiv:2309.10808*, 2023.

[6] T. Kurihana, E. Moyer, R. Willett, D. Gilton, and I. Foster, "Data-driven cloud clustering via a rotationally invariant autoencoder," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–25, 2021.

[7] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, *et al.*, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[9] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[11] E. Racah, C. Beckham, T. Maharaj, S. Ebrahimi Kahou, M. Prabhat, and C. Pal, "Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events," *Advances in neural information processing systems*, vol. 30, 2017.

[12] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "Weatherbench: a benchmark data set for data-driven weather forecasting," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, p. e2020MS002203, 2020.

[13] J. Nathaniel, J. Liu, and P. Gentine, "Metaflux: Meta-learning global carbon fluxes from sparse spatiotemporal observations," *Scientific Data*, vol. 10, no. 1, p. 440, 2023.

[14] S. Rasp, M. S. Pritchard, and P. Gentine, "Deep learning to represent subgrid processes in climate models," *Proceedings of the National Academy of Sciences*, vol. 115, no. 39, pp. 9684–9689, 2018.

[15] B. Stocks, J. Mason, J. Todd, E. Bosch, B. Wotton, B. Amiro, M. Flannigan, K. Hirsch, K. Logan, D. Martell, *et al.*, "Large forest fires in canada, 1959–1997," *Journal of Geophysical Research: Atmospheres*, vol. 107, no. D1, pp. FFR–5, 2002.

[16] M.-A. Parisien, V. S. Peters, Y. Wang, J. M. Little, E. M. Bosch, and B. J. Stocks, "Spatial patterns of forest fires in canada, 1980–1999," *International Journal of Wildland Fire*, vol. 15, no. 3, pp. 361–374, 2006.

[17] K. R. Knapp, M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, "The international best track archive for climate stewardship (ibtracs) unifying tropical cyclone data," *Bulletin of the American Meteorological Society*, vol. 91, no. 3, pp. 363–376, 2010.

[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.