TOWARDS AN IMPROVED UNDERSTANDING AND UTILIZATION OF MAXIMUM MANIFOLD CAPACITY REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Maximum Manifold Capacity Representations (MMCR) is a recent multi-view self-supervised learning (MVSSL) method that matches or surpasses other leading MVSSL methods. MMCR is interesting because it does not fit neatly into any of the commonplace MVSSL families, instead originating from a statistical mechanical perspective on the linear separability of data manifolds. We seek to better understand and then better utilize MMCR. To better understand MMCR, we leverage tools from high dimensional probability to demonstrate that MMCR incentivizes alignment and uniformity of learned embeddings. We then leverage tools from information theory to show that such embeddings maximize a well-known lower bound on mutual information between views, thereby connecting the geometric perspective of MMCR to the information-theoretic perspective often discussed in MVSSL. To better utilize MMCR, we mathematically predict and experimentally confirm non-monotonic changes in the pretraining loss akin to double descent but with respect to atypical hyperparameters. We also discover compute scaling laws that enable predicting the pretraining loss as a function of gradients steps, batch size, embedding dimension and number of views. We then show that MMCR, originally applied to image data, is performant on multimodal image-text data. Broadly, by more deeply understanding the theoretical and empirical behavior of MMCR, our work reveals powerful insights on improving MVSSL methods.

1 INTRODUCTION

Yerxa et al. (2023) recently proposed a new MVSSL method named MMCR that achieves superior-to-similar performance than leading MVSSL methods. For background on MVSSL, see App. A. MMCR is interesting for at least two reasons. Firstly, MMCR does not fit neatly into any MVSSL family: it is not contrastive, it performs no clustering, it leverages no distillation, and it does not reduce redundancy. Secondly, unlike many MVSSL methods that originate in information theory, MMCR's foundation lies in the statistical mechanical characterization of the linear separability of data manifolds. In this work, we seek to better understand MMCR and utilize this understanding to drive implementation decisions. Our contributions are summarized in App. B.

2 PRELIMINARIES

Multi-View Self-Supervised Learning (MVSSL) Let $f_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ denote a neural network with parameters θ. Suppose we have a dataset of P points $\{\boldsymbol{x}_p\}_{p=1}^P$ and a set of random transformations (augmentations) $\mathcal{T}$. For each datum $\boldsymbol{x}_p$ in a batch of inputs, we sample K transformations $t^{(1)}, t^{(2)}, ..., t^{(K)} \sim \mathcal{T}$ yielding a set of augmented views: $v^{(1)}(\boldsymbol{x}_p), ..., v^{(K)}(\boldsymbol{x}_p)$. We feed these transformed data into the network and obtain *embeddings* Z:

$$\boldsymbol{z}_p^{(k)} \overset{\text{def}}{=} f_\theta(t^{(k)}(\boldsymbol{x}_p)) \in \mathcal{Z}.$$

In practice, $\mathcal{Z}$ is commonly the D-dimensional hypersphere $\mathbb{S}^{D-1} \overset{\text{def}}{=} \{\boldsymbol{z} \in \mathbb{R}^D : \boldsymbol{z}^T \boldsymbol{z} = 1\}$ or $\mathbb{R}^D$. Given that we will later touch on information theory, we need notation to refer to the random variables; we use $Z_p^{(k)}$ to denote the random variable for the embedding whose realization is $\boldsymbol{z}_p^{(k)}$.

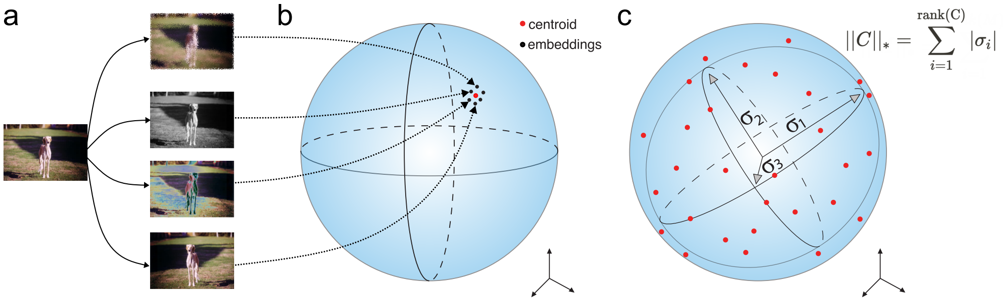

Figure 1: **Schematic of Maximum Manifold Capacity Representations (MMCR).** Left: $K \geq 2$ views are generated of each datum, then embedded through a deep neural network on the surface of the hypersphere. Center: For each datum, the *centroid* of the embeddings is computed. Right: The MMCR pretraining loss minimizes negative nuclear norm of the centers is then minimized.

Maximum Manifold Capacity Representations Maximum Manifold Capacity Representations (MMCR) (Yerxa et al., 2023) originates from classical results regarding performance of linear binary classifiers (Cover, 1965; Gardner, 1987; 1988; Chung et al., 2018). MMCR proceeds in the following manner: MMCR takes the embeddings output by the network and normalizes them to lie on the hypersphere: $\boldsymbol{z}_p^{(1)}, ..., \boldsymbol{z}_p^{(K)} \in \mathbb{S}^{D-1}$. Then, MMCR computes the *center* (average) of the embeddings for each datum: $\boldsymbol{c}_p \overset{\text{def}}{=} \frac{1}{K}\sum_k \boldsymbol{z}_p^{(k)}$. Next, MMCR forms a $P \times D$ matrix C where the n-th row of C is the center $\boldsymbol{c}_p$ and defines the loss:

$$\mathcal{L}_{MMCR} \overset{\text{def}}{=} -\|C\|_* \overset{\text{def}}{=} -\sum_{r=1}^{rank(C)} \sigma_r(C),$$

where $\sigma_r(C)$ is the r-th singular value of C and $\|\cdot\|_*$ is the nuclear (trace, Schatten 1) norm .

3 A High-Dimensional Probability Analysis of MMCR

We consider MMCR's regime of large number of patterns P and high embedding dimension D and show that the MMCR loss $\mathcal{L}_{MMCR}$ can be minimized by (a) making each center $\boldsymbol{c}_p = \frac{1}{K}\sum_k \boldsymbol{z}_p^{(k)}$ lie on the surface of the hypersphere, and (b) making the distribution of centers as close to uniform on the hypersphere as possible. We begin by adopting two useful definitions from prior works (Wang & Isola, 2020; Gálvez et al., 2023):

Definition 3.1 (Perfect Reconstruction)**.** We say a network f_θ achieves *perfect reconstruction* if $\forall \boldsymbol{x} \in \mathcal{X}, \forall\, t^{(1)}, t^{(2)} \in \mathcal{T}, \boldsymbol{z}^{(1)} = f_\theta(t^{(1)}(\boldsymbol{x})) = f_\theta(t^{(2)}(\boldsymbol{x})) = \boldsymbol{z}^{(2)}$.

Definition 3.2 (Perfect Uniformity)**.** Let $p(Z)$ be the distribution over the network representations induced by the data sampling and transformation sampling distributions. We say a network f_θ achieves *perfect uniformity* if the distribution $p(Z)$ is the uniform distribution on the hypersphere.

We will show that a network that achieves both perfect reconstruction and perfect uniformity obtains the lowest possible MMCR loss by first showing that $\mathcal{L}_{MMCR}$ has a lower bound and then showing that such a network achieves this bound.

Proposition 3.3. *Suppose that* $\forall p \in [P], \boldsymbol{c}_p^T \boldsymbol{c}_p \leq 1$. *Then,* $0 \leq ||C||_* \leq \sqrt{P \min(P, D)}$.

Proof. App. C. □

Proposition 3.4. *Let f_θ achieve perfect reconstruction. Then,* $\|\boldsymbol{c}_p\|_2 = 1\ \forall n$.

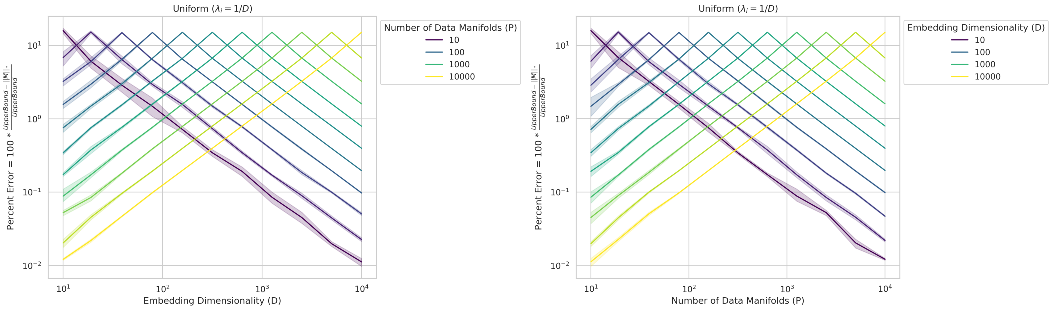

Figure 2: **Embeddings with perfect reconstruction and perfect uniformity achieve the lowest possible MMCR loss.** Away from the $P = D$ threshold, uniform random vectors achieve the theoretically derived upper bound on the nuclear norm of the mean matrix M i.e. the lower bound on $\mathcal{L}_{MMCR}$. The gap between the network's loss and the lowest possible $\mathcal{L}_{MMCR}$ falls $\propto P^{-1}$ (left) or $\propto D^{-1}$ (right) away from the $P = D$ threshold.

Proof. Because f_θ achieves perfect reconstruction, $\forall n, \forall t^{(1)}, t^{(2)}$, $\boldsymbol{z}_p^{(1)} = \boldsymbol{z}_p^{(2)}$. Thus $\boldsymbol{c}_p = (1/K) \sum_k \boldsymbol{z}_p^{(k)} = (1/K) \sum_k \boldsymbol{z}_p^{(1)} = \boldsymbol{z}_p^{(1)}$, and since $\|\boldsymbol{z}_p^{(1)}\|_2 = 1$, we have $\|\boldsymbol{c}_p\|_2 = 1$. □

Theorem 3.5. *Let $f_\theta : \mathcal{X} \to \mathbb{S}^D$ be a network that achieves perfect reconstruction and perfect uniformity. Then f_θ achieves the lower bound of $\mathcal{L}_{MMCR}$ with high probability. Specifically:*

$$\|C\|_* = \begin{cases} P(1 - O(P/D)) & \text{if } P \leq D \\ \sqrt{PD}(1 - O(D/P)) & \text{if } P \geq D \end{cases}$$

with high probability in $\min(P, D)$.

Proof. App. D. □

4 An Information-Theoretic Understanding of MMCR

Many MVSSL methods originate in information theory or can be understood from an information theoretic perspective (Oord et al., 2018; Bachman et al., 2019; Wang & Isola, 2020; Wu et al., 2020; Gálvez et al., 2023; Shwartz-Ziv et al., 2023). Based on our newfound understanding of what distributions of embeddings MMCR incentivizes, how can we connect MMCR's statistical mechanical geometric viewpoint to an information theoretic viewpoint? Consider the mutual information between the embeddings of two different views $Z^{(1)}$ and $Z^{(2)}$ of some input datum. The mutual information between the two views must be at least as great as the sum of two terms:

$$I[Z^{(1)}; Z^{(2)}] \geq \underbrace{\mathbb{E}_{p(Z^{(1)}, Z^{(2)})}[\log q(Z^{(1)}|Z^{(2)})]}_{\text{Reconstruction}} + \underbrace{H[Z^{(1)}]}_{\text{Entropy}}, \tag{1}$$

where $q(Z^{(1)}|Z^{(2)})$ is a variational distribution because the true distribution $p(Z^{(1)}|Z^{(2)})$ is unknown. This bound is well-known, e.g., (Cover, 1965; Wang & Isola, 2020; Gálvez et al., 2023), but we repeat them to show how MMCR connects to an information-theoretic perspective.

Theorem 4.1. *Let $f_\theta : \mathcal{X} \to \mathbb{S}^D$ be a network, and let the number of views per datum be constant. Let $\mathcal{Q}$ be the variational family of distributions on the hypersphere. Then f_θ maximizes the mutual information lower bound (Eqn. 1) iff f_θ achieves perfect reconstruction and perfect uniformity.*

Proof. Perfect reconstruction maximizes reconstruction. Perfect uniformity maximizes entropy. □

Thus, a minimizer of MMCR is a maximizer of this mutual information lower bound.

Theorem 4.2. *Let f_{θ^*} be a network that achieves perfect reconstruction and perfect uniformity, let the number of views per datum be constant, and let $\mathcal{Q}$ be the variational family of distributions on the hypersphere. Then f_{θ^*} is both a minimizer of $\mathcal{L}_{MMCR}$ and a maximizer of the variational lower bound of mutual information Eqn. 1.*

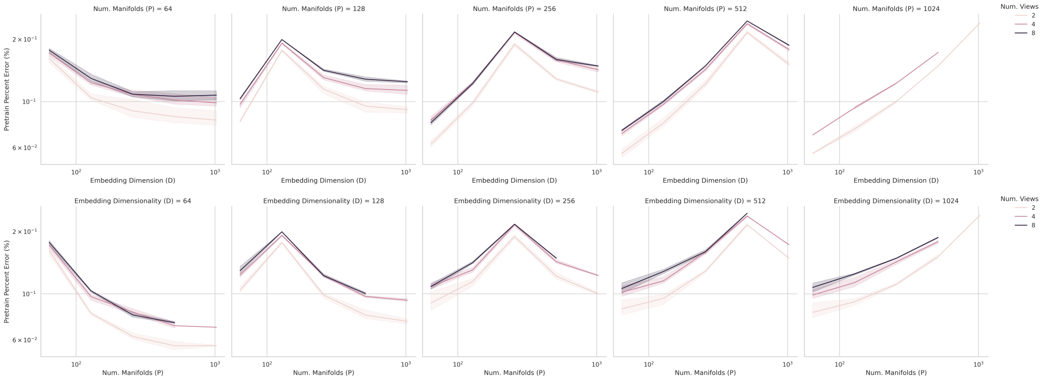

Figure 3: **Double-Descent in Maximum Manifold Capacity Representations.** As predicted mathematically, MMCR's pretraining percent error $\overset{\text{def}}{=} (\sqrt{P \min(P, D)} - ||C||_*)/\sqrt{P \min(P, D)}$ exhibits non-monotonic double descent-like behavior, peaking when the number of data points P equals the number of dimensions D. On either side of the $P = D$ threshold, the pretraining percent error falls. Networks are ResNet-18s pretrained on STL-10's "unlabeled" split.

Proof. Theorem 3.5 and Theorem 4.1. □

5 Double Descent in MMCR Pretraining Loss

An unexpected and interesting insight from our high-dimensional probability analysis (Theorem 3.5) is a prediction that the Maximum Manifold Capacity Representations (MMCR) pretraining loss should also exhibit a non-monotonic double descent-like behavior in its pretraining loss. Double descent is a well-known machine learning phenomenon where the test loss exhibits non-monotonic changes as a function of the total number of data and the number of model parameters; see App. F for citations. However, our analysis suggests that this double descent-like behavior should occur with respect to atypical parameters: the number of manifolds P and the number of dimensions D, rather than the number of data and the number of model parameters. Specifically, our theory predicts suggests that the highest pretraining error should occur exactly at the threshold $P = D$, with pretraining error falling on either side of the threshold.

$$\text{Pretraining Percent Error}(C) \overset{\text{def}}{=} \frac{\sqrt{P \min(P, D)} - ||C||_*}{\sqrt{P \min(P, D)}}$$

We pretrained ResNet-18s (He et al., 2016) on STL-10 (Coates et al., 2011), a dataset similar to CIFAR-10 but higher resolution (96x96x3) and containing an additional unlabeled split of 100000 images. We swept $P \in \{64, 128, 256, 512, 1024\} \times D \in \{64, 128, 256, 512, 1024\} \times K \in \{2, 4, 8\}$, where K is the number of views. For all combinations of number of points P, number of dimensions D and number of views K, we found that the pretraining percent error peaked when $P = D$ (Fig. 3) and declined on either side of the $P = D$ threshold.

6 Compute Scaling Laws in MMCR

In many MVSSL methods, changing hyperparameters often renders the pretraining losses incommensurate, making comparisons between runs difficult if not impossible. However, because the MMCR pretraining percent error yields a quantity bounded between 0 and 1, we can compare different training runs with different hyperparameter values for the number of data points P and data dimensionality D. Performing such a comparison yields a second interesting empirical phenomenon: compute neural scaling laws in the MMCR pretraining percent error. See App. G.

REFERENCES

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2019.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections, 2023.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=xm6YD62D1Ub`.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, jan 2020. doi: 10.1137/20m1336072. URL `https://doi.org/10.1137%2F20m1336072`.

Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, pp. 517–526. PMLR, 2017.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

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

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

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning, 2022.

SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

Aidan Clark, Diego De Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International Conference on Machine Learning*, pp. 4057–4086. PMLR, 2022.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. A u-turn on double descent: Rethinking parameter counting in statistical learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Robert PW Duin. Classifiers in almost empty spaces. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pp. 1–7. IEEE, 2000.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

Borja Rodrıguez Gálvez, Arno Blaas, Pau Rodriguez, Adam Golinski, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. The role of entropy and reconstruction in multi-view self-supervised learning. In *International Conference on Machine Learning*, pp. 29143–29160. PMLR, 2023.

Elizabeth Gardner. Maximum storage capacity in neural networks. *Europhysics letters*, 4(4):481, 1987.

Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, 2021.

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.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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*, pp. 9729–9738, 2020.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Andy L Jones. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Anders Krogh and John A Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135, 1992.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Oren Neumann and Claudius Gros. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Manfred Opper. Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pp. 922–925, 1995.

Tomaso Poggio, Gil Kur, and Andrzej Banburski. Double descent in the condition number. *arXiv preprint arXiv:1912.06190*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Jason W Rocks and Pankaj Mehta. The geometry of over-parameterized regression and adversarial perturbations. *arXiv preprint arXiv:2103.14108*, 2021.

Jason W Rocks and Pankaj Mehta. Bias-variance decomposition of overparameterized regression with random linear features. *Physical Review E*, 106(2):025304, 2022a.

Jason W Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical Review Research*, 4(1):013201, 2022b.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2019.

Rylan Schaeffer, Mikail Khona, Zachary Robertson, Akhilan Boopathy, Kateryna Pistunova, Jason W. Rocks, Ila Rani Fiete, and Oluwasanmi Koyejo. Double descent demystified: Identifying, interpreting and ablating the sources of a deep learning puzzle, 2023a.

Rylan Schaeffer, Zachary Robertson, Akhilan Boopathy, Mikail Khona, Ila Fiete, Andrey Gromov, and Sanmi Koyejo. Divergence at the interpolation threshold: Identifying, interpreting & ablating the sources of a deep learning puzzle. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023b.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim GJ Rudner, and Yann LeCun. An information-theoretic perspective on variance-invariance-covariance regularization. *arXiv preprint arXiv:2303.00633*, 2023.

Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Yao-Hung Hubert Tsai, Martin Q. Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=068E_JSq9O`.

F Vallet. The hebb rule for learning linearly separable boolean functions: learning and generalization. *Europhysics Letters*, 8(8):747, 1989.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Roman Vershynin, Y Eldar, and Gitta Kutyniok. Compressed sensing, theory and applications. In *Introduction to the non-asymptotic analysis of random matrices*, pp. 210–268. Cambridge Univ. Press, 2012.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020.

Thomas Yerxa, Yilun Kuang, Eero Simoncelli, and SueYeon Chung. Learning efficient coding of natural images with maximum manifold capacity representations. *arXiv preprint arXiv:2303.03307*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Jiachen Zhu, Rafael M Moraes, Serkan Karakulak, Vlad Sobol, Alfredo Canziani, and Yann LeCun. Tico: Transformation invariance and covariance contrast for self-supervised visual representation learning. *arXiv preprint arXiv:2206.10698*, 2022.

Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6002–6012, 2019.

A Background on Multi-View Self-Supervised Learning

Multi-View Self-Supervised Learning (MVSSL; also known as Joint-Embedding Self-Supervised Learning) is a powerful approach to unsupervised learning. The idea is to create multiple transformations, or "views", of unsupervised data, then use these views in a supervised-like manner to learn generally useful representations. MVSSL methods are diverse but can be loosely grouped into different families (Balestriero et al., 2023): (1) contrastive, e.g., CPC (Oord et al., 2018), MoCo 1 (He et al., 2020), SimCLR (Chen et al., 2020a), MoCo 2 (Chen et al., 2020b), CMC (Tian et al., 2020), RPC (Tsai et al., 2021) and TiCo (Zhu et al., 2022); (2) clustering e.g., Noise-as-Targets (Bojanowski & Joulin, 2017), DeepCluster (Caron et al., 2018), Self-Labeling (Asano et al., 2019), Local Aggregation (Zhuang et al., 2019), SwAV (Caron et al., 2020); (3) distillation/momentum e.g., BYOL (Grill et al., 2020), DINO (Caron et al., 2021), SimSiam (Chen & He, 2021); and (4) redundancy reduction e.g., Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2022), TiCo (Zhu et al., 2022). Many MVSSL methods either explicitly originate from information theory (Oord et al., 2018; Bachman et al., 2019) or can be understood from an information-theoretic perspective (Wang & Isola, 2020; Wu et al., 2020; Gálvez et al., 2023; Shwartz-Ziv et al., 2023).

B Contributions

Our contributions are:

1. We leverage tools from high dimensional probability to show that embeddings with perfect invariance and perfect uniformity minimize the MMCR pretraining loss with high probability. This analysis involves bounding the MMCR pretraining loss, allowing us to define a "pretraining percent error" for MMCR; this pretraining percent error then reveals two interesting empirical phenomena (below).
2. We connect this distribution of embeddings to information theory by showing that such a distribution maximizes a well-known variational lower bound on the mutual information between embeddings of multiple views.
3. Our analysis of the MMCR pretraining loss predicts a double descent-like behavior in the pretraining percent error as a function of two parameters: the number of manifolds N and the embedding dimensionality D. We empirically test and confirm this prediction in ResNet-18s He et al. (2016) pretrained on STL-10 Coates et al. (2011). This is notable because (to the best of our knowledge) double descent has not been observed in MVSSL and because these parameters differ from the typical double descent parameters (number of data and number of model parameters).
4. Our pretraining percent error additionally enables comparing different hyperparameters on the MMCR pretraining loss – an ability not commonly available in MVSSL methods – which reveals the existence of compute scaling laws.
5. We demonstrate that MMCR, originally proposed purely for images, can be similarly performant in the multi-modal image+text setting. We show that MMCR applied to image+text pairs can match CLIP Radford et al. (2021) on DataComp Small containing 128M high quality image+caption pairs Gadre et al. (2023).

C Proof of Prop 3.3

Proof. Let $\sigma_1, \dots, \sigma_{\min(P,D)}$ denote the singular values of C, so that $\|C\|_* = \sum_{i=1}^{\min(P,D)} \sigma_i$. The lower bound follows by the fact that singular values are nonnegative. For the upper bound, we have

$$\sum_{i=1}^{\min(P,D)} \sigma_i^2 = \mathrm{Tr}\big[CC^T\big] = \sum_{n=1}^{P} \boldsymbol{c}_p^T \boldsymbol{c}_p \leq P.$$

Then, by Cauchy-Schwarz on the sequences $(1, \ldots, 1)$ and $\big(\sigma_1, \ldots, \sigma_{\min(P,D)}\big)$, we get

$$\begin{aligned}\sum_{i=1}^{\min(P,D)} \sigma_i &\leq \sqrt{\left(\sum_{i=1}^{\min(P,D)} 1\right)\left(\sum_{i=1}^{\min(P,D)} \sigma_i^2\right)} \\ &\leq \sqrt{\min(P,D)\,P}.\end{aligned}$$

□

D Proof of Theorem 3.5

Recall that $\mathcal{L}_{MMCR} = -\|C\|_*$ is minimized when $\|C\|_*$ is maximized and that $\|C\|_*$ is upper bounded by $\sqrt{ND}$ if $N > D$ and N if $N < D$ (Proposition 3.3). We want to show a network that achieves perfect reconstruction and perfect uniformity achieves this upper bound on the nuclear norm (equivalently, lower bound on the MMCR loss).

Following the proof of Proposition 3.3, let $\sigma_1, \ldots, \sigma_{\min(N,D)}$ denote the singular values of C, so that $\|C\|_* = \sum_i \sigma_i$. By Proposition 3.4, we have

$$\sum_i \sigma_i^2 = \mathrm{Tr}\big[CC^T\big] = \sum_{n=1}^{N} \boldsymbol{\mu}_n^T \boldsymbol{\mu}_n = N.$$

Now, by the equality version of Cauchy–Schwarz on the sequences $(1, \ldots, 1)$ and $\big(\sigma_1, \ldots, \sigma_{\min(N,D)}\big)$, we have

$$\sum_i \sigma_i = \sqrt{\min(N,D)\left(\sum_i \sigma_i^2 - \sum_i \left(\sigma_i - \frac{\sum_j \sigma_j}{\min(N,D)}\right)^2\right)}. \tag{2}$$

So if we can bound this "variance" of the singular values $\sum_i \left(\sigma_i - \frac{\sum_j \sigma_j}{\min(N,D)}\right)^2$, we can show that $\|C\|_*$ closely matches the upper bound obtained in Proposition 3.3.

To do this, let us consider matrix $\sqrt{D}C$. The vectors $\boldsymbol{\mu}_n$ are uniform over the D-dimensional hypersphere $\mathbb{S}^D$, so its rows $\sqrt{D}\boldsymbol{\mu}_n$ have mean zero, are isotropic, and (by Example 5.25 in Vershynin et al. (2012)) are sub-gaussian with parameter $\|\sqrt{D}\boldsymbol{\mu}_n\|_{\psi_2} = O(1)$.[1] Therefore,

- **If $\mathbf{N \leq D}$**, then (using the fact that $\|\boldsymbol{\mu}_n\|_2 = 1$ for all $n \in [N]$) we can to apply Theorem 5.58 in Vershynin et al. (2012) on the transpose of $\sqrt{D}C$, obtaining that for any $t \geq 0$, the singular values of $\sqrt{D}C$ are within $\sqrt{D} \pm O(\sqrt{N}) + t$ with probability at least $1 - 2\exp(-\Omega(t^2))$. Setting t to a large enough multiple of $\sqrt{N}$, they are all within $\sqrt{D} \pm O(\sqrt{N})$ with probability at least $1 - 2\exp(-N)$. Consequently, with the same probability, the singular values of C are all within $\pm O(\sqrt{N/D})$ of each other, and we get $\sum_i \left(\sigma_i - \frac{\sum_j \sigma_j}{\min(N,D)}\right)^2 \leq N \cdot O\left(\sqrt{N/D}\right)^2 = O(N^2/D)$. Plugging this into Eqn. 2, we get $\|C\|_* \leq \sqrt{N(N - O(N^2/D))} = \sqrt{N}(1 - O(N/D))$.
- **If $\mathbf{N \geq D}$**, then we can apply Theorem 5.39 in Vershynin et al. (2012) on $\sqrt{D}C$, obtaining that for any $t \geq 0$, the singular values of $\sqrt{D}C$ are within $\sqrt{N} \pm O(\sqrt{D}) + t$ with probability at least $1 - 2\exp(-\Omega(t^2))$. Setting t to a large enough multiple of $\sqrt{D}$, they are all within $\sqrt{N} \pm O(\sqrt{D})$ with probability at least $1 - 2\exp(-D)$. Consequently, with the same probability, the singular values of C are all within $\pm O(1)$ of each other, and we get $\sum_i \left(\sigma_i - \frac{\sum_j \sigma_j}{\min(N,D)}\right)^2 \leq D \cdot O(1)^2 = O(D)$. Plugging this into Eqn. 2, we get $\|C\|_* \leq \sqrt{D(N - O(D))} = \sqrt{ND}(1 - O(D/N))$.

[1] Here, $\|\cdot\|_{\psi_2}$ denotes the sub-gaussian norm (intuitively, the "effective standard deviation" of a sub-gaussian random variable). For a scalar random variable X, it is defined as $\|X\|_{\psi_2} := \sup_{p \geq 1} p^{-1/2}(\mathbb{E}[|X|^p])^{1/p}$ (Definition 5.7 in Vershynin et al. (2012)), and for a random vector $\boldsymbol{u} \in \mathbb{R}^D$, it is defined as $\|\boldsymbol{u}\|_{\psi_2} := \sup_{\boldsymbol{v} \in \mathbb{S}^D} \|\boldsymbol{u}^T\boldsymbol{v}\|_{\psi_2}$ (Definition 5.22 in Vershynin et al. (2012)).

E Python Code for Perfect Reconstruction and Perfect Uniformity Embeddings

To test our claim that networks which achieve perfect reconstruction and perfect uniformity achieve the nuclear norm upper bound, we sample a uniform distribution of centroids (thereby enforcing reconstruction by construction) and measure the nuclear norm relative to our claimed upper bound. Python code for our simulations is included below:

```
import pandas as pd
import numpy as np


N_list = np.logspace(start=1, stop=4, num=11).astype(int)
D_list = np.logspace(start=1, stop=4, num=11).astype(int)
repeats = np.arange(5).astype(int)
uniform_distribution_nuclear_norm_data_list = []

for N in N_list:
    for D in D_list:
        print(f”N:■{N}\tD:■{D}”)
        for repeat in repeats:
            embeddings = np.random.normal(loc=0, scale=10.0, size=(N, D))
            embeddings /= np.linalg.norm(embeddings, axis=1, keepdims=True)
            row = {
                ”Spectrum”: ”uniform”,
                ”Number■of■Data■Manifolds■(N)”: N,
                ”Embedding■Dimensionality■(D)”: D,
                ”Repeat”: repeat,
                ”Nuclear■Norm”: np.linalg.norm(embeddings, ord=”nuc”),
            }
            uniform_distribution_nuclear_norm_data_list.append(row)

uniform_distribution_nuclear_norm_df = pd.DataFrame(
    uniform_distribution_nuclear_norm_data_list
)
```

F Double Descent

Double descent citations: Vallet (1989); Krogh & Hertz (1991); Geman et al. (1992); Krogh & Hertz (1992); Opper (1995); Duin (2000); Spigler et al. (2018); Belkin et al. (2019); Bartlett et al. (2020); Belkin et al. (2020); Nakkiran et al. (2021); Poggio et al. (2019); Advani et al. (2020); Liang & Rakhlin (2020); Adlam & Pennington (2020); Rocks & Mehta (2022b; 2021; 2022a); Mei & Montanari (2022); Hastie et al. (2022); Bach (2023); Schaeffer et al. (2023a); Curth et al. (2023); Schaeffer et al. (2023b)

G Compute Scaling Laws in MMCR

Scaling laws are another wide-spread phenomenon of interest in machine learning where the pretraining loss follows a predictable power law-like trend with respect to specific quantities such as number of parameters, total number of data or amount of compute (typically measured in floating point operations) (Hestness et al., 2017; Rosenfeld et al., 2019; Henighan et al., 2020; Kaplan et al., 2020; Gordon et al., 2021; Hernandez et al., 2021; Jones, 2021; Zhai et al., 2022; Hoffmann et al., 2022; Clark et al., 2022; Neumann & Gros, 2022).

By plotting the ResNet-18 networks pretrained on STL-10, once can clearly see power law scaling in the pretraining percent error with the amount of compute (floating point operations) for all number of points P, embedding dimensions D, and number of views K (Fig. 4). A key detail is that

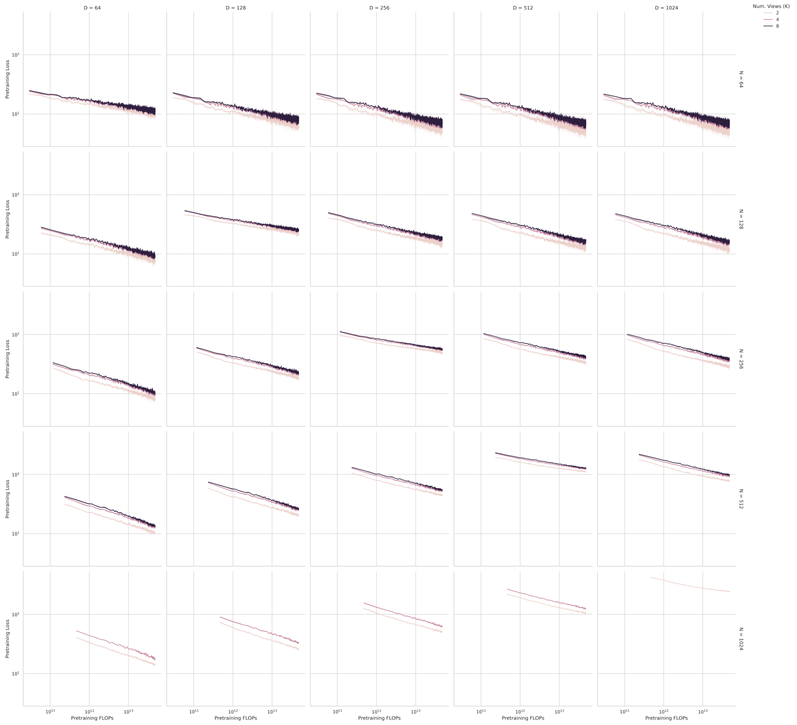

Figure 4: **Compute Scaling Laws.** For all values of number of points P (equivalently, batch size), embedding dimension D and number of views per datum K, the pretraining percent error falls predictably as a power law with the amount of compute i.e. total floating point operations. Consistent with the double descent-like findings in Fig. 3, the on-diagonal subfigures (corresponding to $P = D$) exhibit higher initial pretraining percent errors and less steep slopes with compute than the off-diagonal subfigures (corresponding to $P \neq D$).

these neural scaling curves highlight the double descent-like behavior: the on-diagonal subfigures (corresponding to runs where $P = D$) have both higher pretraining percent error and a less sleep slope for the pretraining percent error, meaning that the pretraining percent error starts higher and falls more slowly. The takeway is that practictioners would be well advised to not pretrain networks where the number of points P (i.e. the batch size) equals the embedding dimension D.

H Multi-Modality in MMCR

We next demonstrate that MMCR can be high-performing in a decidedly more challenging setting: multimodal self-supervised learning. Specifically, we consider the setting of OpenAI's Contrastive Language-Image Pretraining model (CLIP) Radford et al. (2021), in which two different networks are pretrained on image-text caption pairs.

In this multimodal setting, two networks f_θ and $g_{\theta'}$ embed data from two different data domains X and Y. X and Y are paired, such that every example in X has a corresponding positive pair in Y and vice versa. As such, from an MMCR perspective, X and Y can be understood as two "views" of the same underlying object. The optimal transformed embeddings $f_\theta(X)$ and $g_{\theta'}(Y)$ therefore should map to the same space, and we can use our improved understanding of MMCR to train these optimal networks.

The notable difference between this setting and the commonplace MVSSL setting is first that X and Y might represent extremely different distributions in practice, and second f_θ and g_θ are two separate and different neural network architectures. CLIP is a prominent example of such a cross-modal feature alignment task between a text encoder and an image encoder Radford et al. (2021). In

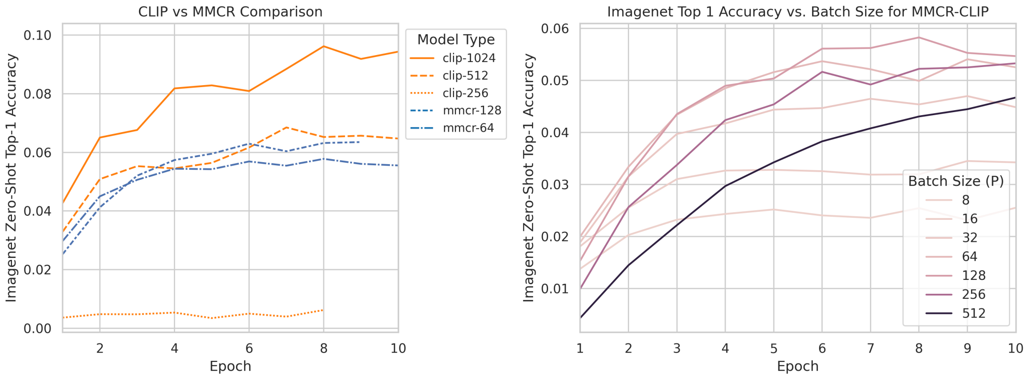

Figure 5: **Multimodal MMCR on Image-Text Caption Pairs.** Left: Multimodal MMCR vs Contrastive Language-Image Pretraining (CLIP) performance on ImageNet measured in zero-shot top-1 accuracy. Multimodal MMCR outperforms CLIP for smaller batch sizes but underperforms CLIP for larger batch sizes. Right: Imagenet top-1 accuracy sweep over batch sizes for MMCR. Unlike CLIP, MMCR exhibits non-monotonic performance scaling with batch size, and best results are found at intermediate batch sizes. To generate strong validation performance scaling behavior, MMCR requires that both batch size and dimension increase simultaneously.

this paper, we investigate whether applying the MMCR objective to the CLIP setting can improve the quality of learned representations.

In image-text alignment, we have access to image-text pairs, which are respectively fed through a vision encoder (here, a ResNet-50) and a text encoder (a transformer Vaswani et al. (2017)). We apply the MMCR objective between the embeddings produced by the two modalities.

We base our Multimodal MMCR experiments off of the open-source CLIP implementation OpenCLIP Cherti et al. (2022). We apply Multimodal MMCR to DataComp-Small and compare zero-shot Imagenet performance with the standard CLIP objective, which is equivalent to SimCLR with $\tau = 1$. DataComp-Small is the smallest version of the curated DataComp dataset family for training CLIP-style models Gadre et al. (2023). This dataset consists of 128 million high-quality image and text pairs that can be used in multimodal training.

We found that convergence of MMCR in the image-text mapping setting is highly dependent on learning rate, and models will fail to converge for learning rates above $\approx 1e-4$. For all runs, we set our Multimodal MMCR learning rate as $1e-4$ and our normal CLIP learning rate as $1e-3$. With the standard CLIP embedding size of $D = 1024$, we swept performance of our models over the critical hyperparameter of batch size (N), finding the optimal batch size to be 128. We compare the performance of the optimal batch size Multimodal MMCR to normal CLIP (Fig. 5). We find that while Multimodal MMCR outperforms CLIP at small batch sizes (< 512) and remains competitive with CLIP with a batch size of 512, it is underperforms CLIP at higher batch sizes. The CLIP loss is a batch contrastive method, and thus benefits directly from increasing batch size. MMCR, however, is simultaneously batch and dimension contrastive, and as a result to achieve similar scaling it is likely Multimodal MMCR would need to increase the size of its latent embedding space beyond 1024 Garrido et al. (2023).

I Relationship of MMCR to the Duality of Sample-Contrastive and Dimension-Contrastive Self-Supervised Learning

In their ICLR 2023 paper "On the Duality Between Contrastive and Non-Contrastive Self-Supervised Learning", Garrido et al. (2023) noted that contrastive (also known as sample-contrastive) and non-contrastive (also known as dimension-contrastive) SSL methods can be seen as two sides of the same coin. Specifically, letting $Z \in \mathbb{R}^{PK \times D}$ denote the matrix of stacked embeddings, then sample-contrastive methods (e.g., SimCLR) incentivize entropy via:

$$\mathcal{L}_{\text{Sample-Contrastive}} \overset{\text{def}}{=} ||ZZ^T - \text{diag}(ZZ^T)||_F^2,$$

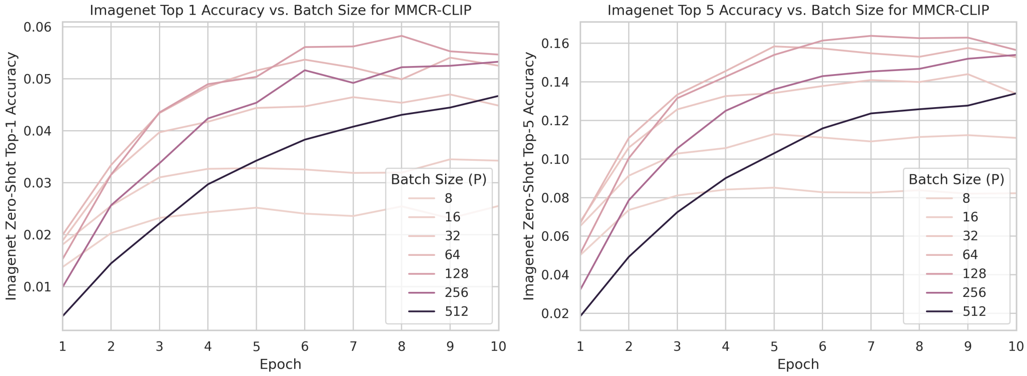

Figure 6: Multimodal MMCR exhibits strong batch size dependence in ImageNet zero-shot validation performance. Intermediate batch sizes, such as 128 and 256, achieved the best validation performance by a large margin. By contrast, the smallest batch sizes or batch sizes closest to the embedding dimension size of 1024 fared the poorest.

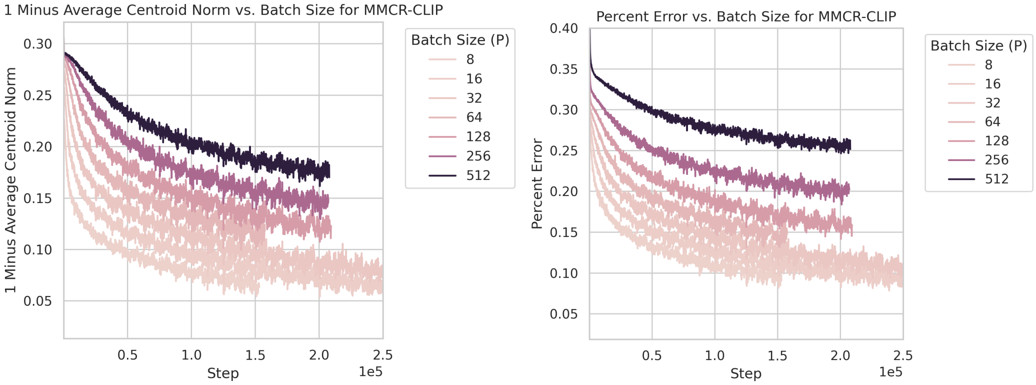

Figure 7: To understand the perplexing batch size dependence, we analyze the complement of the average centroid norm $(1 - \|\mu\|_2^2)$ and the pretaining percent error relative to the lower bound as defined earlier$(1 - \frac{\|M\|_*}{N})$. The complement of the average centroid norm is an unbiased estimator for perfect reconstruction in our network. We find that lower batch sizes converge closer to perfect reconstruction and to lower percent error.

whereas dimension-contrastive methods (e.g., BarlowTwins) incentivize entropy via:

$$\mathcal{L}_{\text{Dimension-Contrastive}} \overset{\text{def}}{=} ||Z^T Z - \text{diag}(Z^T Z)||_F^2.$$

Both families also include an invariant loss $\mathcal{L}_{\text{Invariance}}$ as part of the total loss, typically the MSE between the positive pairs. We observe that both families of loss aim to maximize on-diagonal elements through their invariance loss and minimize off-diagonal elements through their contrastive losses, on a batch-wise or dimension-wise correlation matrix respectively. In both cases, the loss functions aim to maximize the spectra of their matrices (either ZZ^T or $Z^T Z$), and given that these matrices have related spectra, one might wonder why not maximize the spectrum of Z directly? Maximizing the spectrum of Z is qualitatively what MMCR aims to do via its nuclear norm-based loss.