

LEARNING COARSE-GRAINED REPRESENTATIONS: AN EXPLORATION OF MUTUAL INFORMATION VIA HYPERSPHERICAL DENSITY

Anonymous authors

Paper under double-blind review

ABSTRACT

We revisit InfoMax for representation learning, using hyperspherical geometry with a non-parametric von Mises-Fisher kernel density estimator and differential entropy. This method is minimal with no asymmetry and trains stably. Results are competitive on smaller datasets such as CIFAR-10, STL-10 and LC25000, but lags behind modern baselines on ImageNet-1000. Experiments show that weakening the global entropy term consistently helps classification accuracy, suggesting that strict mutual information classification favors coarse grouping over fine discrimination.

1 INTRODUCTION

Self-supervised methods keep improving, but much of the progress has come from heuristics: architectural asymmetries, stop-grads, whitening, variance penalties, and carefully tuned augmentations. These choices work, yet the connection to first principles is often indirect. In this work, we revisit the InfoMax principle (Linsker, 1988b), and study a different estimator: a hyperspherical formulation with a non-parametric density and differential entropy. Our goal is to test how this theoretical backed formulation behaves, study what it is actually learning and where it breaks compared to modern SSL methods.

In this paper, we study an implementation of the InfoMax principle, while keeping it close to the mathematical foundation. We evaluate it with multiple datasets, showing that while it works well in smaller datasets, the learned representations end up representing coarser structures as the number of classes increases. We show that the low accuracy in ImageNet-1000 is the result of these coarser structures, not a fundamental failure in the training process. Additionally, we performed several experiments to determine how certain hyperparameters affect the classification accuracy, and found that weakening the global separation makes accuracy improvement slower, but leads to higher accuracy given enough epochs, while a stronger global separation leads to earlier higher accuracy but eventual learning stagnation.

2 RELATED WORK

The InfoMax principle goes back to early neural coding work, where maximizing output entropy under simple constraints was proposed as a driver for useful features (Linsker, 1988a). It later grounded classic ICA via entropy maximization at the output layer (Bell & Sejnowski, 1995). A complementary line formalizes representation learning through explicit mutual information terms and variational bounds (Barber & Agakov, 2003; Belghazi et al., 2018).

Several deep methods are framed around maximizing mutual information between inputs and learned representations. Deep InfoMax pursues both global and local agreements (Hjelm et al., 2019). AMDIM extends this with multi-scale and patch-level objectives (Bachman et al., 2019). Contrastive predictive coding casts InfoMax through a classification view and the InfoNCE bound (van den Oord et al., 2019), a perspective that underlies later contrastive systems (Chen et al., 2020; He et al., 2020). A different but related path maximizes MI between cluster assignments for segmentation/cluster discovery (Ji et al., 2019).

Beyond InfoNCE, there are broader MI estimators and f-divergence bounds used in practice (Barber & Agakov, 2003; Belghazi et al., 2018; Nguyen et al., 2007). Work in information-theoretic learning also explores entropy and divergence criteria directly (Principe, 2010). Our approach sits in this space: we use a minimal, simple estimator on the hypersphere with differential entropy, keeping the objective close to first principles.

A parallel theoretical foundation argues for shaping information rather than maximizing it: preserve what predicts the task while discarding nuisance variability (Tishby et al., 2000). Many recent SSL objectives can be read as mixing an InfoMax drive (agreement across views) with implicit bottlenecks or regularizers—whitening, variance penalties, stop-grad, or architectural asymmetries (Zbontar et al., 2021; Bardes et al., 2021; Grill et al., 2020). Our study focuses on the pure InfoMax side and documents where that alone succeeds (coarse structure) and where extra bias appears necessary (fine-grained recognition).

3 APPROACH

Our method is based on the InfoMax principle. The objective of InfoMax is to maximize the mutual information between the source and the learned representations. The mutual information is defined as $I(X; Y) = H(Y) - H(Y|X)$. Calculating entropy in high-dimensional spaces is usually considered intractable. For that reason, we opted to restrict the representation geometry to a hypersphere. Under that geometry, we can use the von Mises-Fisher (vMF) kernel to estimate the density in a non-parametric way.

For unit vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{S}^{D-1}$ with cosine similarity $s_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$, the von Mises-Fisher (vMF) kernel with concentration κ and normalization constant $C_D(\kappa)$ is:

$$K_\kappa(\mathbf{x}_i, \mathbf{x}_j) = C_D(\kappa) \exp(\kappa s_{ij}), \quad (1)$$

Then, we estimate the global differential entropy as:

$$H_{\text{global}} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^N K_\kappa(\mathbf{x}_i, \mathbf{x}_j) \right) \quad (2)$$

and the local differential entropy (only between the set $\mathcal{P}(i)$ of positive pairs of i) as:

$$H_{\text{local}} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{1}{N-1} \sum_{j \in \mathcal{P}(i)} K_\kappa(\mathbf{x}_i, \mathbf{x}_j) \right) \quad (3)$$

Finally, after adding two α and β factors (default to 1), we can express the mutual information formula as

$$I(X; Y) \approx \alpha \cdot H_{\text{global}} - \beta \cdot H_{\text{local}} \quad (4)$$

For generating the views of each sample, we opted to a similar augmentation recipe as DINO (Caron et al., 2021). In our case, we used two global views and six local views. Each global view takes a crop of 40% to 100% of the sample. And each local view takes a crop of 5% to 40% of the sample.

Unless mentioned otherwise, all trainings were done using AdamW optimizer, a batch size of 128 and a learning rate of $1e-3$. To make our runs reproducible, we implemented our code using the Lightly Framework and used the evaluation setup developed by Kalapos & Gyires-Tóth (2024). Models trained on LC25000 were trained for 25000 epochs, the ones trained on ImageNet-1000 were trained for 100 epochs, and the ones trained on CIFAR-10 and STL-10 were trained for 200 epochs.

4 MAIN RESULTS

We trained and evaluated our method on CIFAR-10 (Krizhevsky & Hinton, 2009), STL-10 (Coates et al., 2011), LC-25000 (Borkowski et al., 2019) and ImageNet-1000 (Deng et al., 2009) datasets.

Dataset	Method	Arch	Linear		K-NN	
			Top 1	Top 5	Top 1	Top 5
STL-10	Supervised	RN18	71.86	97.41	72.43	87.69
STL-10	<i>Ours</i>	RN18	81.98	99.25	76.85	93.83
STL-10	<i>Ours</i>	VIT-T16	85.98	99.66	80.31	94.65
STL-10	VICReg	RN18	81.69	99.14	78.48	93.85
STL-10	SimCLR	RN18	82.89	99.36	79.16	94.91
STL-10	Barlow Twins	RN18	82.19	99.04	78.20	93.98
STL-10	BYOL	RN18	81.28	99.28	77.14	94.39
STL-10	DINO	RN18	79.75	99.11	76.85	93.53
CIFAR-10	Supervised	RN18	84.17	98.97	84.16	93.36
CIFAR-10	<i>Ours</i>	RN18	72.54	97.97	68.11	89.14
CIFAR-10	VICReg	RN18	74.72	98.14	72.67	90.59
CIFAR-10	SimCLR	RN18	72.58	97.86	69.16	90.25
CIFAR-10	Barlow Twins	RN18	74.63	97.91	71.81	90.87
CIFAR-10	BYOL	RN18	69.37	97.81	65.54	89.74
CIFAR-10	DINO	RN18	72.46	98.02	70.07	90.34
LC-25000	Supervised	RN18	99.83	-	100.00	-
LC-25000	<i>Ours</i>	RN18	99.75	-	99.99	-
LC-25000	<i>Ours</i>	VIT-T16	98.54	-	99.89	-
LC-25000	VICReg	RN18	99.30	-	100.00	-
LC-25000	SimCLR	RN18	98.35	-	99.99	-
LC-25000	Barlow Twins	RN18	98.09	-	99.98	-
LC-25000	BYOL	RN18	95.94	-	99.54	-
LC-25000	DINO	RN18	98.46	-	99.99	-
IM-1000	<i>Ours</i>	RN50	52.12	77.81	35.37	64.50
IM-100*	<i>Ours*</i>	RN50	68.15	89.68	55.25	77.38

Table 1: Training results. Linear and K-NN accuracies (top-1 and top-5). *IM-100 metrics were calculated using the ImageNet-1000 trained model.

The network architectures trained were ResNet-18 (He et al., 2016), ResNet-50 (for ImageNet-1000) and ViT-Tiny (Dosovitskiy et al., 2021) (for STL-10 and LC-25000). The training was done on a single NVIDIA H100 GPU with 80GB of GPU memory hosted in the cloud. The classification accuracy is summarized in Table 1. While the results were good for datasets with a low number of “latent classes”, it is clear that our method did not fare well for more diverse datasets such as ImageNet-1000. However, the gap between top 1 and top 5 accuracy suggests that despite the low top 1 accuracy, our method still produced a good coarse structure. We can confirm this if we take our ImageNet-1000 model and evaluate its linear probe and K-NN accuracy on the ImageNet-100 dataset: the results are much better than in ImageNet-1000, even if the gap is smaller.

4.1 EMERGENT FOCUSING ON OBJECTS

Neural networks trained using our method learned to focus on main objects in an image. This behavior was detected both in the ResNet50 trained on ImageNet-1000 and a ViT-Tiny trained on STL-10. As shown in Figure 1, when visualized the gradients using SmoothGrad-CAM technique applied to the ResNet-50 trained on ImageNet-1000, we can see that the network learned with some success to focus on the main objects in each image. On the other hand, in Figure 2, we visualized the average CLS-to-patch attention across all transformers blocks of a ViT-Tiny trained on STL-10 using our method. In both cases, we can see that the networks learned to focus on the main objects in the image.

5 IMPACT OF HYPERPARAMETERS

We explored variations of the number of dimensions in the projector, the impact of different κ (bandwidth parameter) in the final accuracy and the weight of the local entropy estimation (β parameter).

Unlike other methods such as Barlow Twins, we did not find any noticeable difference for changing the number of dimensions of the projector. On the other hand, κ influence is more noticeable.

5.1 PROJECTOR DIMENSIONS

We trained a ResNet-18 using our method on STL-10, changing the number of dimensions of the final projector to 128, 256, 512, 1024, 2048 and 4096. There is no clear advantage of increasing the size of the final projector, and the learning curves are very similar in all cases, as shown in Figure 3.

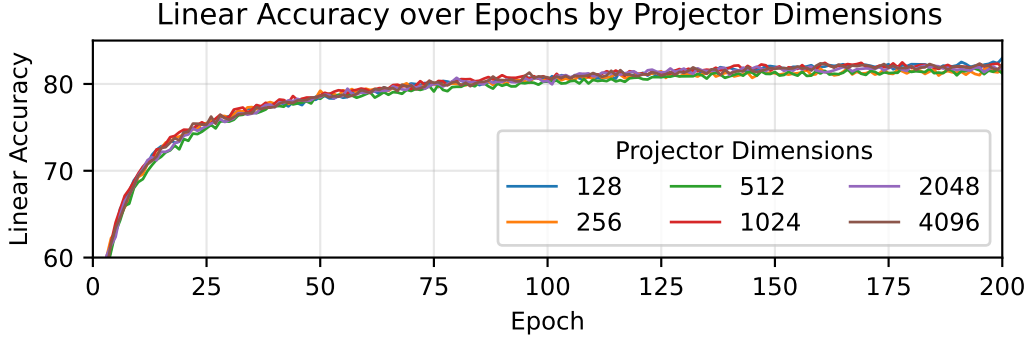


Figure 3: Linear probe top 1 accuracy per epoch, varying the number of dimensions of the projector.

5.2 BANDWIDTH PARAMETER

We trained a ResNet-18 using our method on STL-10, varying the bandwidth parameter κ to 0.1, $0.6931 \approx \ln 2$, 1, 10 and 20. Recall that κ is the concentration (bandwidth) of the vMF kernel used in our non-parametric density estimate: larger κ produces a sharper kernel and a smaller κ a smoother one. In this case, we found that higher κ values accelerate early accuracy, however, if it is too high it hinders the learning process after the initial epochs. Smaller κ values lead to slower accuracy improvements, but given enough epochs seems to eventually catch up. In the Figure 4 the accuracy per epoch is shown for each κ evaluated.

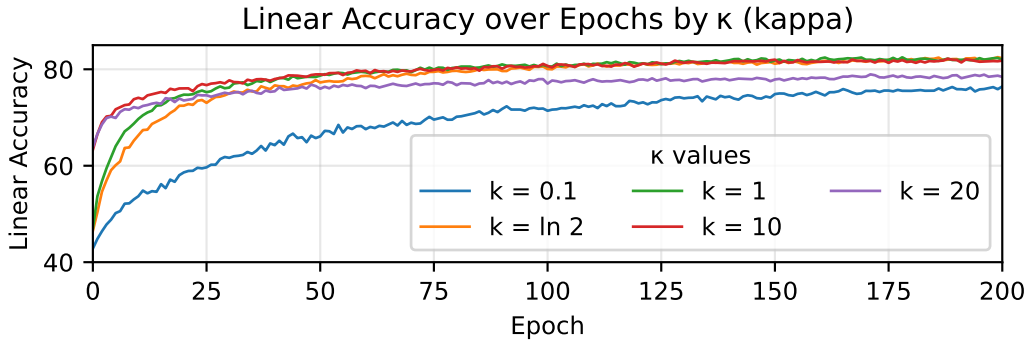


Figure 4: Linear probe top 1 accuracy per epoch, varying the bandwidth parameter.

5.3 LOCAL ENTROPY ESTIMATION WEIGHT

Additionally, we modified the weight of the local entropy estimation. We found that reducing the β weight, alternatively weighting separation more, leads to higher early accuracy. However, these setups seems to get stuck after some epochs, and runs with higher local weight end up having a higher final accuracy, despite the slower early accuracy gains. The accuracy during training of those variations is shown in Figure 5. Below in that figure, a zoom in section is shown, where we can

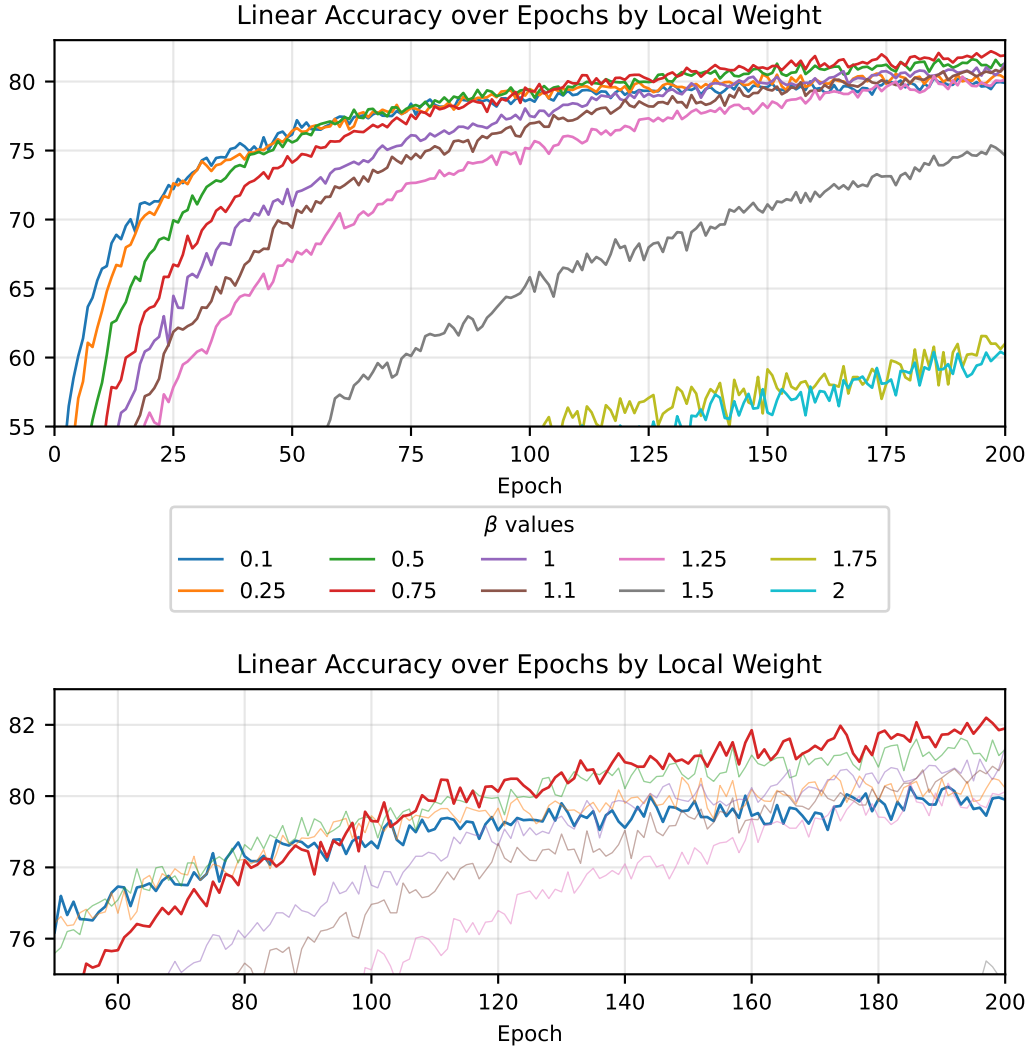


Figure 5: Linear accuracy over epochs by changing the β weight (above). Below, a zoom in, highlighting the curves for $\beta = 0.1$ and $\beta = 0.75$

appreciate how the accuracy of $\beta = 0.1$ stalls compared to other settings, particularly $\beta = 0.75$. This suggests that longer training would benefit from a softer separation force. Which is interesting, as this suggests that at least for this sort of classification problems, the best representation may not be the one with maximum mutual information.

6 CONCLUSION

In this paper, we discussed an implementation of the InfoMax idea of maximizing Mutual Information. In this case, we opted to use hypersphere geometry, non-parametric kernel density estimators and differential Shannon entropy. We found that this method can generate good representations in datasets with low number of classes, but it starts to lag behind for datasets with higher level of classes. However, the metrics suggest that lower accuracy is due to a learned coarse structure, and not a failure to learn good representations. Our experiments additionally point in the direction that we can get better representations if we relax our global entropy term. This suggests that, at least for classification, representations should not look to maximize mutual information, but instead there

may be another theoretical objective that may explain better how to produce strong fine-grained structures.

REFERENCES

- Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf.
- David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems, NIPS'03*, pp. 201–208, Cambridge, MA, USA, 2003. MIT Press.
- Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.
- Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/belghazi18a.html>.
- Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, November 1995. ISSN 1530-888X. doi: 10.1162/neco.1995.7.6.1129. URL <http://dx.doi.org/10.1162/neco.1995.7.6.1129>.
- Andrew A. Borkowski, Marilyn M. Bui, L. Brannon Thomas, Catherine P. Wilson, Lauren A. DeLand, and Stephen M. Mastorides. Lung and colon cancer histopathological image dataset (lc25000), 2019.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML'20*. JMLR.org, 2020.
- Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL <https://proceedings.mlr.press/v15/coates11a.html>.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 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.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.
- R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL <https://openreview.net/forum?id=Bklr3j0cKX>.
- Xu Ji, Andrea Vedaldi, and Joao Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9864–9873, 2019. doi: 10.1109/ICCV.2019.00996.
- András Kalapos and Bálint Gyires-Tóth. Whitening consistently improves self-supervised learning, 2024. URL <https://arxiv.org/abs/2408.07519>.
- Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988a. doi: 10.1109/2.36.
- Ralph Linsker. An application of the principle of maximum information preservation to linear systems. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988b. URL https://proceedings.neurips.cc/paper_files/paper/1988/file/ec8956637a99787bd197eacd77acce5e-Paper.pdf.
- XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/72da7fd6d1302c0a159f6436d01e9eb0-Paper.pdf.
- Jose C. Principe. *Information Theoretic Learning: Renyi’s Entropy and Kernel Perspectives*. Springer New York, 2010. ISBN 9781441915702. doi: 10.1007/978-1-4419-1570-2. URL <http://dx.doi.org/10.1007/978-1-4419-1570-2>.
- Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL <https://arxiv.org/abs/physics/0004057>.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL <https://arxiv.org/abs/1807.03748>.
- 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.

A LEARNING CURVES

The curves for linear top-1 accuracy are shown at Figure 6. It is interesting to note that no method surpassed CIFAR-10 supervised classification, while in STL-10 case, supervised classification got stuck around 70% linear top-1 accuracy.

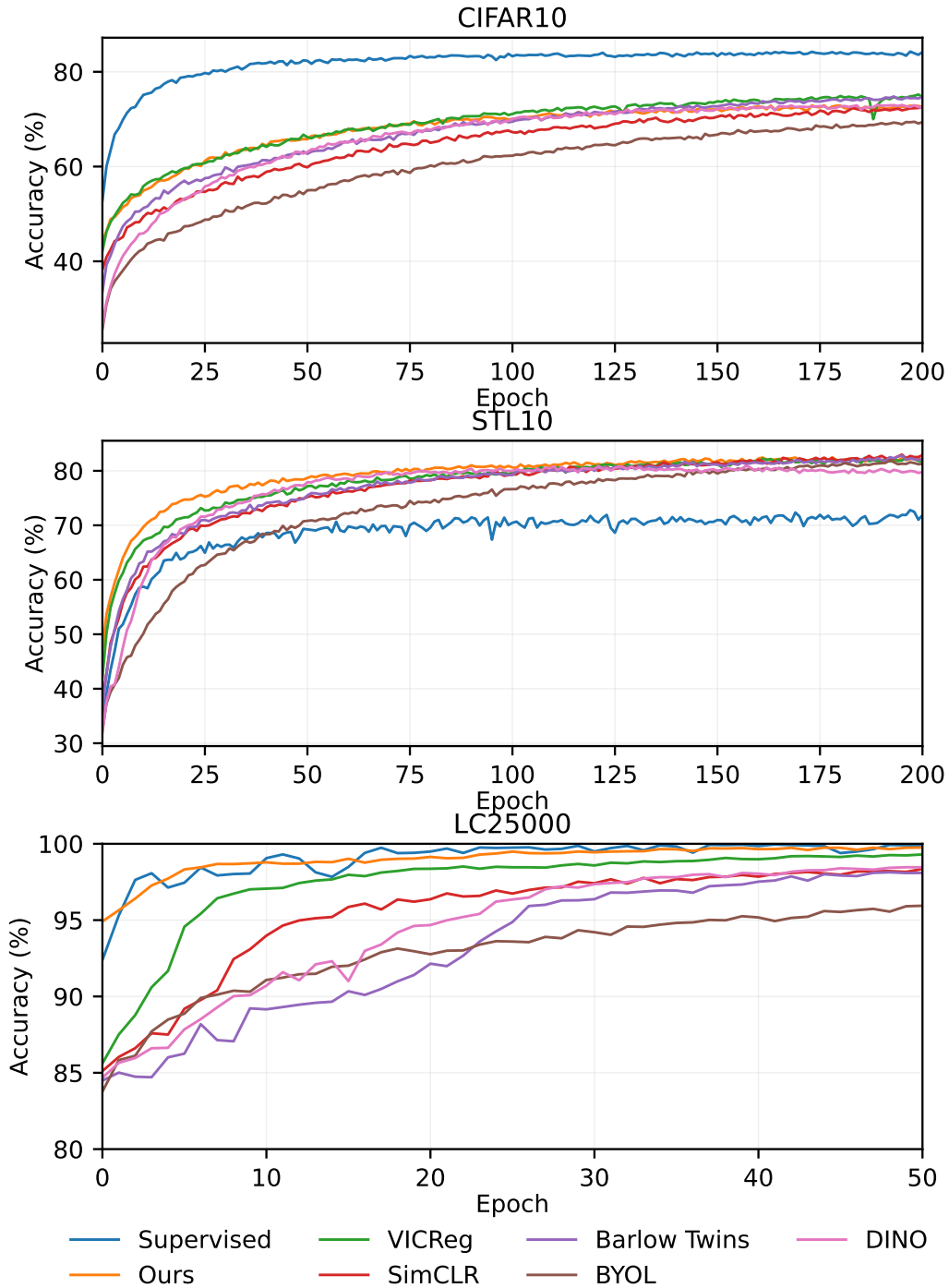


Figure 6: Top 1 accuracy of a linear classifier trained at the end of each epoch.

B IMAGENET-100 CLASS CENTROID SIMILARITY

To visualize how the learned representations may overlap between classes, we opted to create a heatmap of the cosine similarity of the class centroids of ImageNet-100 classes, shown in Figure 7. For this, we used a ResNet-50 trained on ImageNet-1000. From this visualization, we can infer that the model has some classes with more overlap than others. For instance, the big square in the middle corresponds to animals. That square itself can be divided in dogs and other animals, being the dogs the most intense square in the lower right corner, and other four legged wild animals above at the left. The rest of the animals are in the weaker big square at the right bottom of the visualization. The weak square at the top left is dominated by human created objects.

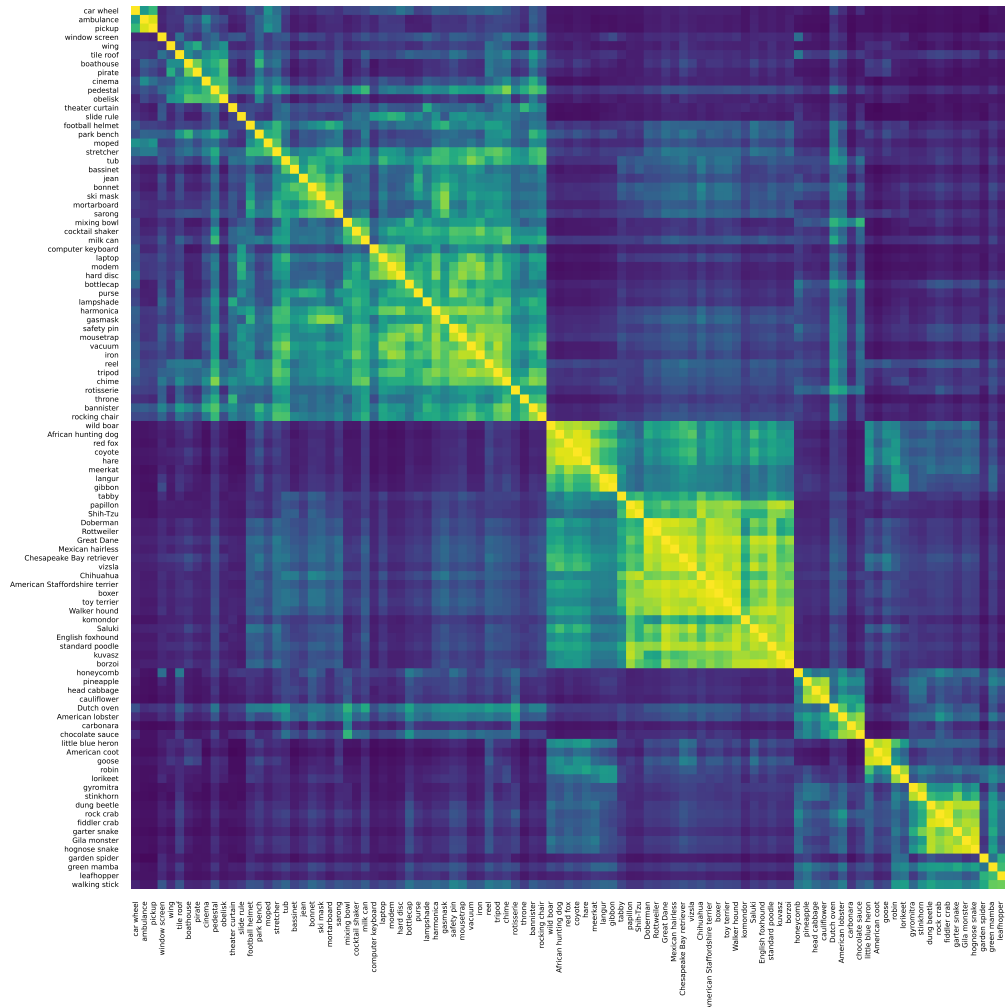


Figure 7: Cosine similarity of ImageNet-100 class centroids, calculated using a ResNet-50 trained on ImageNet-1000.