Simplicial Embeddings in Self-Supervised Learning and Downstream Classification

Samuel Lavoie[◇†], **Christos Tsirigotis**[◇†], **Max Schwarzer**[◇†], **Ankit Vani**[◇†], **Michael Noukhovitch**[◇†], **Kenji Kawaguchi**[‡], **Aaron Courville**[◇†♣]
[◇] Mila, [†] Université de Montréal, [‡] National University of Singapore, [♣] CIFAR Fellow
{samuel.lavoie.m,mnoukhov,aaron.courville}@gmail.com
{christos.tsirigotis,max.schwarzer,ankit.vani}@umontreal.ca
kenji@comp.nus.edu.sg

Abstract

Simplicial Embeddings (SEM) are representations learned through self-supervised learning (SSL), wherein a representation is projected into L simplices of V dimensions each using a `softmax` operation. This procedure conditions the representation onto a constrained space during pretraining and imparts an inductive bias for group sparsity. For downstream classification, we formally prove that the SEM representation leads to better generalization than an unnormalized representation. Furthermore, we empirically demonstrate that SSL methods trained with SEMs have improved generalization on natural image datasets such as CIFAR-100 and ImageNet. Finally, when used in a downstream classification task, we show that SEM features exhibit emergent semantic coherence where small groups of learned features are distinctly predictive of semantically-relevant classes.

1 Introduction

Over-complete representations are representations of an input that are non-unique combinations of a number of basis vectors greater than the input's dimensionality [Lewicki and Sejnowski, 2000]. Mostly studied in the context of the sparse-coding literature [Gregor and LeCun, 2010; Goodfellow et al., 2012; Olshausen, 2013], sparse over-complete representations have been shown to increase stability in the presence of noise [Donoho et al., 2006], have applications in neuroscience [Olshausen and Field, 1996; Lee et al., 2007] and lead to more interpretable representations [Murphy et al., 2012; Fyshe et al., 2015; Faruqui et al., 2015]. However, the choice of basis vectors is generally assumed to be learned using traditional methods such as ICA [Teh et al., 2003] or fitting linear models [Lewicki and Sejnowski, 2000], limiting the expressive power of the encoding function.

Meanwhile, self-supervised learning (SSL) is an emerging family of methods that aims to learn an encoding of the data without manual supervision, such as through class labels, and using neural network encoders. Recent work [Hjelm et al., 2019; Grill et al., 2020; Saeed et al., 2020; You et al., 2020] learn dense representations that can solve complex tasks by simply fitting a linear model on top of the learned representation. While this demonstrates SSL's efficacy, we demonstrate that sparse and overcomplete representations can further improve the downstream performance of these methods.

Inspired by work on language emergence, Dessì et al. [2021] propose to induce a discrete representation at the output of the encoder in a SSL model. Contrary to their work, we demonstrate that hard-discretization during pre-training is not necessary to achieve a sparse representation. Instead, we propose to project the encoder's output into L vectors of V dimensions onto which we apply a `softmax` function to impart an inductive bias toward sparse vectors [Correia et al., 2019; Goyal et al., 2022], also alleviating the need to use high-variance estimators to back-propagate the gradient through the encoder. We refer to this embedding as Simplicial Embeddings (SEM) because the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

softmax functions map the unnormalized representations onto L simplices. The procedure to induce SEM is simple, efficient, and generally applicable.

The SSL pre-training phase, used with SEM, learns the set of L *approximately*-sparse vectors. Key to controlling the inductive bias of SEM during pre-training is the softmax temperature parameter: the lower the temperature, the stronger the bias toward sparsity. Consistent with earlier attempts at sparse representation learning [Coates and Ng, 2011], we find that the optimal sparsity for pre-training need not correspond to the optimal for downstream learning.

For downstream classification, we may discretize the learned representation by, for example, taking the argmax for each simplex. But, we can also leverage the SEM to control the representation's expressivity via the softmax's temperature.

Empirically, we provide evidences that SEM is applicable to most recent SSL methods and lead to a better representation. For the seven SSL methods probed [Chen et al., 2020; He et al., 2020; Grill et al., 2020; Caron et al., 2020, 2021; Zbontar et al., 2021; Bardes et al., 2022], SEM increases the accuracy by 2% to 4% on CIFAR-100 over the baseline without SEM. We observe constant improvement as we increase the number of vectors L showing benefits of overcomplete representation in SEM. We also observe important improvement when training a SSL method with SEM on ImageNet on in-distribution test sets as well as several out-of-distribution test sets and transfer learning benchamarks, demonstrating the potential of SEM for large scale applications. Finally, we perform a qualitative analysis, and find that SEM learns features that are closely aligned to the semantic categories extant in the data, demonstrating evidence of disentangled and more interpretable representations, as it was previously observed in overcomplete representations [Faruqui et al., 2015].

2 Simplicial Embeddings

Simplicial Embeddings (SEM) are representations that can be integrated easily into a contrastive learning model [Hjelm et al., 2019; Chen et al., 2020], the BYOL method [Grill et al., 2020], and other SSL methods [Caron et al., 2020, 2021; Zbontar et al., 2021]. For example, in BYOL, we insert SEM after the encoder and before the projector and the rest is unchanged as shown in Figure 2c. There, t and t' are augmentations defined by the practitioner, ξ would be a moving average of θ defined as follow: $\xi \leftarrow \alpha\xi + (1-\alpha)\theta$, with $\alpha \in [0,1]$. The SSL loss is defined as the cosine similarity.

To produce an SEM representation, the encoder's output e_θ is embedded into L vectors $z_i \in \mathbb{R}^V$. A temperature parameter τ scales z_i then a softmax re-normalizes each vector z_i to produce $\bar{z}_i$. Finally, the normalized vectors $\bar{z}_i$ are concatenated to produce the vector $\hat{z}$ of length $L \cdot V$. Formally, the re-normalization is as follows:

$$\bar{z}_i := \sigma_\tau(z_i), \quad \sigma_\tau(z_i)_j = \frac{e^{z_{ij}/\tau}}{\sum_{k=1}^{V} e^{z_{ik}/\tau}}, \quad \hat{z} := \text{Concat}(\bar{z}_1, \dots, \bar{z}_L), \quad \forall i \in [L], \forall j \in [V]. \tag{1}$$

3 Empirical analysis

We empirically study the effect of SEM on the representation of SSL methodsand demonstrate that SEM improves the test set accuracy on CIFAR-100 [Krizhevsky, 2009]. On IMAGENET [Deng et al., 2009], we study the effect of SEM on robustness and transfer learning datasets. Finally, we present evidences that features that result from SEMs appear to be more naturally aligned to the semantic categories found in the data.

Training setup For all experiments, we build off the implementation of the baseline models from the Solo-Learn library [da Costa et al., 2021].

We probe the encoder's output for the baseline methods, as typically done in the litterature. For models with SEM, we probe the SEM. In our experiments, the embedder is a linear layer followed by BatchNorm [Ioffe and Szegedy, 2015]. Unless mentioned otherwise, we use $L = 5000$ and $V = 13$ for the SEM representation. We do not perform any search for the non-SEM hyper-parameters. The SEM Hyper-parameters are selected by using a validation set of 10% of the training set of CIFAR-100 and 10 samples per class for IMAGENET. The test accuracy is obtained by retraining the model with all of the training data using the parameters found with the validation set. We use a batch size of 256, and train models for 200 epochs on IMAGENET and 1000 epochs on CIFAR-100.

3.1 SEM improves on downstream classification

We evaluate the effect of adding SEMs in seven modern SSL approaches. We take standard SimCLR [Chen et al., 2020], MoCo-v2 [He et al., 2020], BYOL [Grill et al., 2020] Barlow-Twins [Zbontar et al., 2021], SwAV [Caron et al., 2020], DINO [Caron et al., 2021] and VicReg [Bardes et al., 2022] models and implement SEM after the encoder. We compare our approach on CIFAR-100 with a ResNet-18 in Table 1. The reported numbers are the means and the standard deviations over 5 random seeds. For every SSL methods, using SEMs improves the baseline methods by a margin of 2% to 4% demonstrating that SEM is a general approach that improves in-distribution generalization for SSL methods.

	SimCLR	MoCo	BYOL	Barlow-Twins	SwAV	DINO	VicReg
Baseline	65.8 ± 0.3	69.3 ± 0.3	70.7 ± 0.2	70.7 ± 0.3	64.6 ± 0.3	66.8 ± 0.3	68.5 ± 0.2
With SEM	$\mathbf{69.5 \pm 0.2}$	$\mathbf{71.0 \pm 0.3}$	$\mathbf{73.9 \pm 0.2}$	$\mathbf{73.0 \pm 0.2}$	$\mathbf{67.7 \pm 0.2}$	$\mathbf{69.2 \pm 0.3}$	$\mathbf{71.4 \pm 0.4}$

Table 1: Linear probe accuracy on CIFAR-100 trained for *1000 epochs* with ResNet-18 encoder. We compare the *test accuracy* of several SSL models with and without SEM. The baseline models are taken from [da Costa et al., 2021]. The SEM normalized output ($\hat{z}_\theta$) is used for the linear probe with SEM. **Boldface** indicates highest accuracy. Green rows indicate a SSL method + SEM.

3.2 SEM improvement on large-scale datasets with ImageNet

We demonstrate that SEM improves the accuracy on large scale datasets, such as IMAGENET. We demonstrate that SEM generally improves the test set accuracy on several robustness test sets, transfer learning datasets and semi-supervised via fine-tuning with 1% and 10% of the data. The embedding is pre-trained for 200 epochs using the BYOL SSL procedure.

	IN	IN-V2	IN-R	IN-C	FLOPs
ResNet50:					
BYOL*	70.6	-	-	-	-
BYOL	71.9	59.2	18.8	39.5	$4.1e9$
BYOL+SEM	**74.1**	**61.2**	**22.1**	**43.4**	$4.7e9$
ResNet50-x2:					
BYOL	74.2	62.1	22.2	47.3	$1.1e10$
BYOL+SEM	**75.9**	**63.7**	**23.5**	**48.8**	$1.2e10$
ResNet50-x4:					
BYOL	75.8	64.0	22.9	49.8	$3.7e10$
BYOL+SEM	**77.2**	**64.8**	**25.1**	**52.0**	$3.8e10$

Table 2: Test accuracies of a linear probe trained with the IMAGENET samples on a pre-trained representation trained for *200 epochs*. * Taken from [Chen and He, 2020]

	FOOD101	C10	C100	SUN	DTD	FLOWER
Linear probe:						
BYOL	74.2	91.8	74.9	60.9	**72.2**	88.9
BYOL+SEM	**74.7**	**93.5**	**78.6**	**62.1**	71.9	**91.5**
Fine-tuned:						
BYOL	83.1	97.2	83.6	59.1	68.8	85.4
BYOL+SEM	**84.7**	97.2	**85.6**	**63.3**	**71.3**	**91.7**

Table 3: Transfer learning accuracy by training a linear probe on a pre-trained representation with IMAGENET for *200 epochs*.

	Top-1		Top-5	
	1%	10%	1%	10%
BYOL	51.6	67.5	78.0	88.9
BYOL+SEM	**56.7**	**69.9**	**81.0**	**90.0**

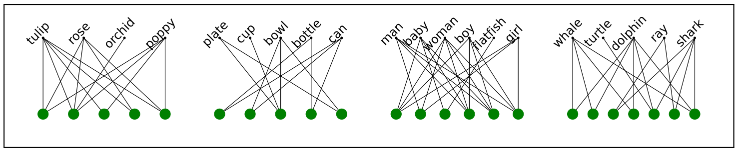
Table 4: Semi-supervised learning top-1 and top-5 accuracy by fine-tuning a model on ImageNet.

3.3 Semantic coherence of SEM features

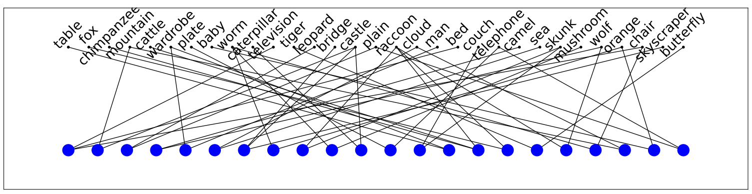
Here we demonstrate that SEM features are coherently aligned with the semantics present in the training data. Qualitatively, we visualize the most predictive features of a downstream linear classifier trained on CIFAR-100 and see that the classes with similar predictive features are semantically related. Quantitatively we propose a metric that returns the ratio of features mostly predictive for a classes that are in the same super class to total number of class predictive for this feature.

For both our analysis, we use a linear classifier trained on the features extracted from BYOL with and without SEM. Consider the trained linear classifier with a weight matrix $W \in \mathbb{R}^{N \times C}$, with N features, and C classes. By preserving the top K parameters of the weight matrix W for each class and pruning the features predictive for only one class, we create a bipartite graph between two set of nodes: the CIFAR-100 classes and the features of the representation. We denote this graph $\mathcal{W}_K$.

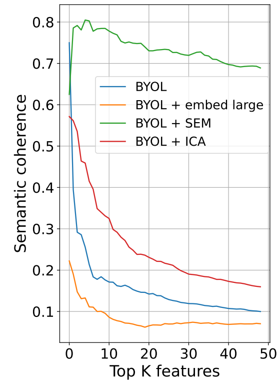
The qualitative analysis is given by plotting the subset $\mathcal{W}_5$, obtained by taking the top 5 features for each class. We present a subset of the graph for BYOL+SEM in Figure 1a and for BYOL in Figure 1b. The full graphs are presented in the Appendix. In the SEM plot, a set of connected components emerge, and the connected components of the graph are semantically related. For example, the



(a) BYOL + SEM



(b) BYOL



(c)

Figure 1: Semantic coherence of the features. **(a)** and **(b)** Subset of $\mathcal{W}_K$, the bipartite graph of the most important features shared between at least two classes of a classifier trained on BYOL + SEM features in **(a)** and BYOL on the encoded features in **(b)**. The connected components emerge without additional interventions in BYOL + SEM. **(c)** Coherence of the top K features to the semantics of the super-class of the categories of CIFAR-100. It is taken as the number of pairwise categories in the same super-class for which a feature is among its top K most predictive features over the total number of pairwise categories.

first set of connected components are flowers, and the last set of connected components are aquatic mammals. The same class coherence is not observed with either the BYOL baseline or with BYOL augmented with a large representation. In particular, we do not see a small number of semantically related connected components. Instead, we see a large fully connected graphs.

Next, we describe how we quantitatively measure the semantic coherence of the features. Notice that two classes share a common predictive feature on $\mathcal{W}_K$ if they are 2-neighbour. Let $\mathcal{N}(c_i)$ returns all pairs (c_i, c_j) for all j 2-neighbour of c_i. Moreover, define the operation is_super(c_i, c_j) which returns 1 if c_i and c_j are from the same CIFAR-100 superclass and 0 otherwise. We reproduce the superclass of CIFAR-100 in Table 7 in the Appendix. We measure semantic coherence as follows:

$$\text{Coherence}(\mathcal{W}_K) := \frac{1}{C} \sum_{i=1}^{C} \frac{\sum_{(c_i, c_j) \in \mathcal{N}(c_i)} \text{is\_super}(c_i, c_j)}{|\mathcal{N}(c_i)|}, \tag{2}$$

where $C = 100$ for CIFAR-100 and $|\cdot|$ is the cardinality of a set.

We compare the semantic coherence of BYOL+SEM with the control experiments on BYOL: regular BYOL, BYOL with an embedding of the same size as BYOL+SEM but without the normalization and BYOL to which we applied linear ICA [Hyvärinen and Oja, 2000] in an attempt to disentangle the features. In Figure 6, we plot the full graph $\mathcal{W}_5$ for BYOL+SEM and the baselines. We observe that using the SEM yields semantically coherent features for all the classes of CIFAR-100. This observation is consistent with the qualitative and quantitative experiments presented earlier and demonstrates that SEM's inductive bias during pre-training leads to features that are semantically coherent with the semantic categories extant in the data.

4 Conclusion

SEMs are representations that can be obtained by embedding partitions of a latent representation using a `softmax` operation. This simple modification leads to improved generalization on downstream classification for several SSL methods. Furthermore, our semantic coherence analysis indicates that SEMs can naturally disentangle the semantic categories of the data without explicit training objectives. We hope that this work motivates the use of SEM with pre-training to learn discrete representation useful for downstream applications and the study of architectural inductive biases for SSL representations towards more explainable and performant models.

Acknowledgments and Disclosure of Funding

The authors are grateful for the insightful discussions with Xavier Bouthillier, Hattie Zhou, Sébastien Lachapelle, Tristan Deleu, Yuchen Lu, Eeshan Dhekane, and Devon Hjelm. We acknowledge funding support from Samsung and Hitachi, as well as support from Aaron Courville's CIFAR CCAI chair. We also wish to acknowledge Mila and Compute Canada for providing the computing infrastructure that enabled this project. Finally, this project would not have been possible without the contribution of the following open source projects: Pytorch [Paszke et al., 2019], Orion [Bouthillier et al., 2022], Solo-Learn [da Costa et al., 2021], Scikit-Learn [Pedregosa et al., 2011], and Numpy [Harris et al., 2020].

References

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`.

Xavier Bouthillier, Christos Tsirigotis, François Corneau-Tremblay, Thomas Schweizer, Lin Dong, Pierre Delaunay, Fabrice Normandin, Mirko Bronzi, Dendi Suhubdy, Reyhane Askari, Michael Noukhovitch, Chao Xue, Satya Ortiz-Gagné, Olivier Breuleux, Arnaud Bergeron, Olexa Bilaniuk, Steven Bocco, Hadrien Bertrand, Guillaume Alain, Dmitriy Serdyuk, Peter Henderson, Pascal Lamblin, and Christopher Beckham. Epistimio/orion: Asynchronous Distributed Hyperparameter Optimization, March 2022. URL `https://doi.org/10.5281/zenodo.3478592`.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *arXiv:2104.14294 [cs]*, May 2021. URL `http://arxiv.org/abs/2104.14294`. arXiv: 2104.14294.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

Adam Coates and Andrew Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, pages 921–928, 2011. URL `https://icml.cc/2011/papers/485_icmlpaper.pdf`.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1223. URL `https://aclanthology.org/D19-1223`.

Victor G. Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. Solo-learn: A library of self-supervised methods for visual representation learning, 2021. URL `https://github.com/vturrisi/solo-learn`.

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*, pages 248–255. Ieee, 2009.

Roberto Dessì, Eugene Kharitonov, and Marco Baroni. Interpretable agent communication from scratch(with a generic visual processor emerging on the side). *CoRR*, abs/2106.04258, 2021. URL `https://arxiv.org/abs/2106.04258`.

D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006. doi: 10.1109/TIT.2005.860430.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1144. URL `https://aclanthology.org/P15-1144`.

Alona Fyshe, Leila Wehbe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. A compositional and interpretable semantic space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 32–41, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1004. URL `https://aclanthology.org/N15-1004`.

Ian J. Goodfellow, Aaron Courville, and Yoshua Bengio. Large-scale feature learning with spike-and-slab sparse coding. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1387–1394, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Anirudh Goyal, Aniket Rajiv Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael Curtis Mozer, and Yoshua Bengio. Coordination among neural modules through a shared global workspace. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=XzTtHjgPDsT`.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 399–406, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf`.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv:1911.05722 [cs]*, March 2020. URL `http://arxiv.org/abs/1911.05722`. arXiv: 1911.05722.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bklr3j0cKX`.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/ioffe15.html`.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=rkE3y85ee`.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area v2. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL `https://proceedings.neurips.cc/paper/2007/file/4daa3db355ef2b0e64b472968cb70f0d-Paper.pdf`.

Michael S. Lewicki and Terrence J. Sejnowski. Learning Overcomplete Representations. *Neural Computation*, 12(2):337–365, 02 2000. ISSN 0899-7667. doi: 10.1162/089976600300015826. URL `https://doi.org/10.1162/089976600300015826`.

Dianbo Liu, Alex M Lamb, Kenji Kawaguchi, Anirudh Goyal ALIAS PARTH GOYAL, Chen Sun, Michael C Mozer, and Yoshua Bengio. Discrete-valued neural communication. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2109–2121. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/10907813b97e249163587e6246612e21-Paper.pdf`.

Brian Murphy, Partha Pratim Talukdar, and Tom Michael Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. In *COLING*, 2012.

B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, June 1996.

Bruno A. Olshausen. Highly overcomplete sparse coding. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder, editors, *Human Vision and Electronic Imaging XVIII*, volume 8651 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 86510S, March 2013. doi: 10.1117/12.2013504.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. *arXiv:1711.00937 [cs]*, May 2018. URL `http://arxiv.org/abs/1711.00937`. arXiv: 1711.00937.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive Learning of General-Purpose Audio Representations. *arXiv:2010.10915 [cs, eess]*, October 2020. URL `http://arxiv.org/abs/2010.10915`. arXiv: 2010.10915.

Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E. Hinton. Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.*, 4(null):1235–1260, dec 2003. ISSN 1532-4435.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer New York, 1996. doi: 10.1007/978-1-4757-2545-2. URL https://doi.org/10.1007%2F978-1-4757-2545-2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. URL http://arxiv.org/abs/1706.03762. arXiv: 1706.03762 version: 5.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to Unseen Domains: A Survey on Domain Generalization. *arXiv:2103.03097 [cs]*, December 2021a. URL http://arxiv.org/abs/2103.03097. arXiv: 2103.03097.

Shulun Wang, Bin Liu, and Feng Liu. Escaping the gradient vanishing: Periodic alternatives of softmax in attention mechanism, 2021b. URL https://arxiv.org/abs/2108.07153.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *CoRR*, abs/2010.13902, 2020. URL https://arxiv.org/abs/2010.13902.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

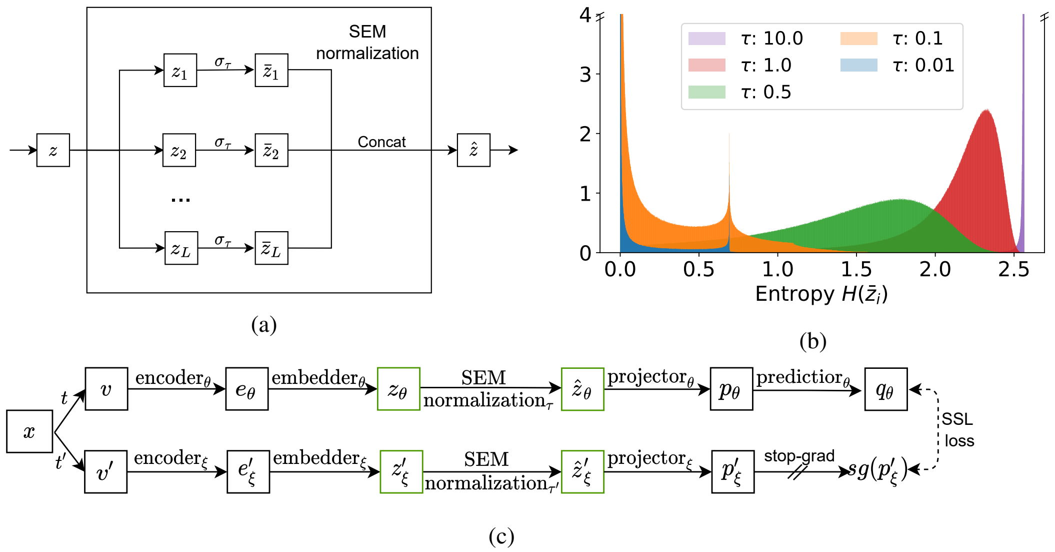


Figure 2: **(a)** Procedure to obtain Simplicial Embeddings (SEM). A matrix $z \in \mathbb{R}^{L \times V}$ contains L vectors $z_i \in \mathbb{R}^V$. The vectors z_i are normalized with σ_τ, the `softmax` operation with temperature τ. The normalized vectors are concatenated into the vector $\hat{z}$. **(b)** Normalized histogram of the entropies $H(\bar{z}_i)$ of each simplex $\bar{z}_i$ for the sample in CIFAR's training dataset at the end of pre-training with various τ. The peak at $\ln(2)$ for $\tau = 0.01$ and $\tau = 0.1$ are a large number of simplices with two elements close to 0.5. **(c)** Integration of SEM with BYOL [Grill et al., 2020]. The encoder outputs a latent vector which is embedded into the matrix $z \in \mathbb{R}^{L \times V}$ and then transformed into SEM.

A Properties of the Simplicial Embeddings

In this subsection, we discuss some of the properties of the Simplcial Embeddings during pre-training and for the downstream tasks.

A.1 Inductive bias towards sparsity during pre-training

In SEM, L controls the numbers of vectors and V controls the number of component that each vectors may have. As such, the higher V is, the sparser the representation is and also the stronger the bias toward sparsity as discussed in Vaswani et al. [2017]; Wang et al. [2021a]. During pre-training, the constraints of the simplex biases each vector towards sparsity by creating a zero-sum competition between the elements. In order for an element of a vector to increase by α, then the other elements must decrease by α and all elements are bounded by 0. For networks to learn useful features, they must prioritize some at the expense of others. For SSL methods with a target network, the temperature for the target network can be different of the online network's as no gradient is back-propagated.

To visualize the effect of the temperature on SEM after pre-training, we interpret each simplex as a probability mass function $p(\bar{z}_{ij})$ where, for all $i \in [L]$, $\sum_{j=1}^{V} p(\bar{z}_{ij}) = 1$ and $p(\bar{z}_{ij}) \geq 0 \, \forall j$. The entropy of a simplex $\bar{z}_i$, defined $H(\bar{z}_i) := -\sum_{j=1}^{V} p(\bar{z}_{ij}) \log p(\bar{z}_{ij})$, informs whether the simplex is a sparse or a dense vector. That is, if $H(\bar{z}_i^{(x)}) = 0$ then the vector is one-hot. On the other hand, if $H(\bar{z}_i^{(x)}) = \ln(V)$ then the vector is dense and uniform. While the temperature τ_p is merely a scaling of the logits, it has an important control over the learned representation's entropy and resulting SEM sparsity. We demonstrate this by learning a representation on CIFAR-100, using BYOL, and analyze the entropies of the resulting simplices. In Figure 2b, we plot the histogram of the entropies $H(\bar{z}_i)$, for a given τ_p, of each simplex for each sample in the training set of CIFAR-100. We observe that, even after pre-training, small temperatures ($\tau_p = 0.01$) yields representations that are close to one-hot vectors while high temperatures yields vectors that are close to uniform vectors.

By pre-training using a `softmax`, SEMs create representations that are conditioned to fit onto simplices. In pre-training, we select τ_p for optimal inductive bias: τ_p too small yields vanishing gradients [Wang et al., 2021b] and τ_p too large yields a bias that is too weak. We may select a different optimal τ_d for downstream performance as discussed formally in the next subsection.

A.2 Reducing the memory footprint of SEM

A large over-complete representation may induce a significant memory footprint due to the additional parameters of the fully connected linear layer used to map to and from the representation. For SEM we require two such mappings as depicted in Figure 2c for BYOL. To reduce the amount of parameters, we propose to partition the matrix multiplication into n small non-overlapping matrix multiplications. Formally, let $v \in \mathbb{R}^{b\times m}$o, $w \in \mathbb{R}^{m\times o}$ and $y = v \cdot w$ be the fully connected matrix multiplication. Instead, we partition v into n blocks with $v^i \in \mathbb{R}^{b\times \frac{m}{n}}$ and define n smaller $w^i \in \mathbb{R}^{\frac{m}{n}\times\frac{o}{n}}$, where $i \in [L]$ is the i^{th} block. Then, we perform a batch matrix multiplication of v^i and w^i that we concatenate as follows: $y^i = v^i \cdot w^i$ and $\bar{y}^i = \text{Concat}([y^1, \dots, y^n])$. Thus, the amount of parameters of this matrix multiplication scales in $\mathcal{O}(\frac{m\cdot o}{n})$, allowing us to reduce the memory consumption by increasing n, the number of blocks.

A.3 SEM improvement on the generalization of the downstream classifier

In this subsection, we mathematically analyze the effect of using SEM for downstream classification. We aim to understand the benefit of training a downstream classifier with SEM normalized input compared to a baseline classifier with unnormalized input. In summary, we show that: (1) there is a trade-off between the training loss and the generalization gap, which is controlled by the value of τ_d, (2) SEM can improve the base model performance when we attain good balance in this trade-off, and (3) the improvement due to SEM is expected to improve or stay constant as L and V increase. In the remaining of this subsection, we consider $\tau = \tau_d$.

Notation We consider a training dataset $S = (z^{(i)}, y^{(i)})_{i=1}^n$ of n samples that is used for supervised training of a classifier using the representation z, which are extracted from the pre-trained model[1], and the corresponding label y. Let g represent the downstream classifier trained on the representation. We can define a baseline model where g is trained without normalization as $f_{\text{base}}(z) = g(z)$ and the corresponding model trained with the SEM normalization of the representation's features as $f_{\text{SEM}\tau}(z) = (g \circ \sigma_\tau)(z)$. Here, σ_τ is applied to each vector z_i. To compare the quality of the base model and the model with SEM normalization, we analyze the generalization gap $\mathbb{E}_{z,y}[l(f(z), y)] - \frac{1}{n}\sum_{i=1}^n l(f(z^{(i)}), y^{(i)})$ for each $f \in \{f^S_{\text{SEM}(\tau)}, f^S_{\text{base}}\}$, where $l : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is the per-sample loss and $f^S_{\text{SEM}(\tau)}$, f^S_{base} are the models obtained from fitting the dataset S.

To simplify the notation, we consider the normalization to $[-1, +1]$; i.e., the encoder's output $z \in \mathcal{Z} = [-1, +1]^{L\times V}$. Next, we define $\mathcal{Q}_i = \{q \in [-1, +1]^V : i = \arg\max_{j\in[V]} q_j\}$, the partition of the space $[-1, +1]^V$. I.e. we have V partitions $\mathcal{Q}_i$ with $i \in [V]$ and $\mathcal{Q}_i$ is the partition of the space where $i = \arg\max_{j\in[V]} q$. We use $\mathcal{Q}_i$ to define the following measure that allows us to understand the effect of the SEM normalization on the representation given to the classifier g and thus to compare f^S_{SEM} and f^S_{base}, a model with and without SEM normalization respectively:

$$\varphi(\sigma_f) = \sup_{i\in[V]} \sup_{q,q'\in\mathcal{Q}_i} \|\sigma_f(q) - \sigma_f(q')\|_2 \tag{3}$$

where $\sigma_{f^S_{\text{SEM}(\tau)}} = \sigma_\tau$ and $\sigma_{f^S_{\text{base}}}$ is the identity function. Here, $\sigma_\tau(q)_j = \frac{e^{q_j/\tau}}{\sum_{t=1}^V e^{q_t/\tau}}$ for $j = 1, \dots, V$. Intuitively, φ is a measure on the expressivity of the representation and depends on V and τ.

We also make the following assumptions. We assume that there exists $\Delta > 0$ such that for any $i \in [L]$, if $k = \arg\max_{j\in[V]} z_{ij}$, then $z_{ik} \geq z_{ij} + \Delta$ for any $j \neq k$. Since Δ can be arbitrarily small (e.g., much smaller than machine precision), this assumption typically holds in practice. Next, we define B to be the upper bound on the per-sample loss such that $l(f(z), y) \leq B$ for all $f \in \mathcal{H}$ and for all $(z, y) \in \mathcal{Z} \times \mathcal{Y}$, where $\mathcal{H}$ is the union of the hypothesis spaces of $f_{\text{SEM}(\tau)}$ and f_{base}. For example, $B = 1$ for the 0-1 loss.

Finally, we define $\mathcal{G}_S$ to be the set of classifiers g returned by the training algorithm using dataset S, and R to be the Lipschitz constant of $l_y \circ g$ for all $y \in \mathcal{Y}$ and $g \in \mathcal{G}_S$; i.e., $|(l_y \circ g)(\sigma_f(z)) - (l_y \circ$

[1]In this subsection, we assume that the extracted representation is z, the embedder's output

$g)(\sigma_f(z'))| \le R\|\sigma_f(z) - \sigma_f(z')\|_F$, where $l_y(g \circ \sigma_f(z)) = l(g \circ \sigma_f(z), y)$. We denote by $c > 0$ a constant in $(n, f, \mathcal{H}, \delta, \mathcal{H}, \tau, S)$.

Using the established notation, Theorem 1 illuminates the advantage of SEM and the effect of the hyper-parameter τ on the performance of the downstream classifier. We present the proof in Appendix B and we present empirical evidence of the theorem's prediction in Figure 5.

Theorem 1. *Let $V \ge 2$. For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for any $f_S \in \{f^S_{\mathrm{SEM}(\tau)}, f^S_{\mathrm{base}}\}$:*

$$\mathbb{E}_{z,y}[l(f_S(z), y)] \le \frac{1}{n}\sum_{i=1}^{n} l(f_S(z^{(i)}), y^{(i)}) + R\sqrt{L\varphi(\sigma_{f_S})} + c\sqrt{\frac{\ln(2/\delta)}{n}}.$$

Moreover,

$$\varphi(\sigma_{f^S_{\mathrm{SEM}(\tau)}}) \to 0 \quad \textit{as } \tau \to 0 \quad \textit{and} \quad \varphi(\sigma_{f^S_{\mathrm{SEM}(\tau)}}) - \varphi(\sigma_{f^S_{\mathrm{base}}}) \le \frac{3}{4}(1 - V) < 0 \quad \forall \tau > 0.$$

The first statement of Theorem 1 shows that the expected loss is bounded by the three terms: the training loss $\frac{1}{n}\sum_{i=1}^{n} l(f_S(z^{(i)}), y^{(i)})$, the second term $R\sqrt{L\varphi(f_S)}$, and the third term $c\sqrt{\frac{\ln(2/\delta)}{n}}$. Since c is a constant in $(n, f, \mathcal{H}, \delta, \mathcal{H}, \tau, S)$, the third term goes to zero as $n \to \infty$ and is the same with and without SEM. Thus, for the purpose of assessing the impact of SEM, we can focus on the second term, where a difference arises.

Theorem 1 shows that $R\sqrt{L\varphi(f_S)}$ goes to zero with SEM; i.e., $\varphi(f^S_{\mathrm{SEM}(\tau)}) \to 0$ as $\tau \to 0$. Also, for any $\tau > 0$, the second term with SEM is strictly smaller than that without SEM as $\varphi(f^S_{\mathrm{SEM}(\tau)}) - \varphi(f^S_{\mathrm{base}}) \le \frac{3}{4}(1 - V) < 0$ and demonstrates that the improvement due to SEM is expected to asymptotically increase as V increases. Also, L is a multiplicative constant of φ which shows that as L increases, the expect improvement due to SEM is also expected to be higher.

Overall, Theorem 1 shows the benefit of SEM as well as the trade-off with τ. When $\tau \to 0$, the second term goes to zero, but the training loss (the first term) can increase due to the reduction in expressivity and increased difficulty in optimization. Thus, τ should be chosen to optimally balance this trade-off.

B Proof of Theorem 1

Let us introduce additional notations used in the proofs. Define $r = (z, y) \in \mathcal{R}$, $\ell(f, r) = l(f(z), y)$,

$$\tilde{\mathcal{C}}_{y,k_1,\ldots,k_L} = \{(z, \hat{y}) \in \mathcal{Z} \times \mathcal{Y} : \hat{y} = y, k_j = \operatorname*{arg\,max}_{t \in [V]} z_{j,t} \quad \forall j \in [L]\},$$

and

$$\tilde{\mathcal{Z}}_{k_1,\ldots,k_L} = \{z \in \mathcal{Z} : k_j = \operatorname*{arg\,max}_{t \in [V]} z_{j,t} \quad \forall j \in [L]\}.$$

We then define $\mathcal{C}_k$ to be the flatten version of $\tilde{\mathcal{C}}_{y,k_1,\ldots,k_L}$; i.e., $\{\mathcal{C}_k\}_{k=1}^{K} = \{\tilde{\mathcal{C}}_{y,k_1,\ldots,k_L,y}\}_{y \in \mathcal{Y}, k_1,\ldots,k_L \in [V]}$ with $C_1 = \tilde{\mathcal{C}}_{1,1,\ldots,1}$, $C_2 = \tilde{\mathcal{C}}_{2,1,\ldots,1}$, $C_{|\mathcal{Y}|} = \tilde{\mathcal{C}}_{|\mathcal{Y}|,1,\ldots,1}$, $C_{|\mathcal{Y}|+1} = \tilde{\mathcal{C}}_{1,2,1,\ldots,1}$, $C_{2|\mathcal{Y}|} = \tilde{\mathcal{C}}_{|\mathcal{Y}|,2,1,\ldots,1}$, and so on. Similarly, define $\mathcal{Z}_k$ to be the flatten version of $\tilde{\mathcal{Z}}_{k_1,\ldots,k_L}$. We also use $\mathcal{Q}_i = \{q \in [-1, +1]^V : i = \arg\max_{j \in [V]} q_j\}$, $\mathcal{I}_k := \mathcal{I}_k^S := \{i \in [n] : r_i \in \mathcal{C}_k\}$, and $\alpha_k(h) := \mathbb{E}_r[\ell(h, r) | r \in \mathcal{C}_k]$. Moreover, we define $\varphi(f^S_{\mathrm{base}}) = \sup_{i \in [V]} \sup_{q,q' \in Q_i} \|q - q'\|_2^2$, and $\varphi(f^S_{\mathrm{SEM}(\tau)}) = \sup_{i \in [V]} \sup_{q,q' \in Q_i} \|\sigma_\tau(q) - \sigma_\tau(q')\|_2^2$ where $\sigma_\tau(q)_j = \frac{e^{q_j/\tau}}{\sum_{t=1}^{V} e^{q_t/\tau}}$ for $j = 1, \ldots, V$.

We first decompose the generalization gap into two terms using the following lemma:

Lemma 1. *For any $\delta > 0$, with probability at least $1 - \delta$,the following holds for all $h \in \mathcal{H}$:*

$$\mathbb{E}_r[\ell(h, r)] - \frac{1}{n}\sum_{i=1}^{n} \ell(h, r_i) \le \frac{1}{n}\sum_{k=1}^{K} |\mathcal{I}_k| \left(\alpha_k(h) - \frac{1}{|\mathcal{I}_k|}\sum_{i \in \mathcal{I}_k} \ell(h, r_i)\right) + c\sqrt{\frac{\ln(2/\delta)}{n}}.$$

Proof. We first write the expected error as the sum of the conditional expected error:

$$\mathbb{E}_r[\ell(h,r)] = \sum_{k=1}^{K} \mathbb{E}_r[\ell(h,r)|r \in \mathcal{C}_k] \Pr(r \in \mathcal{C}_k) = \sum_{k=1}^{K} \mathbb{E}_{r_k}[\ell(h,r_k)] \Pr(r \in \mathcal{C}_k),$$

where r_k is the random variable for the conditional with $r \in \mathcal{C}_k$. Using this, we decompose the generalization error into two terms:

$$\mathbb{E}_r[\ell(h,r)] - \frac{1}{n}\sum_{i=1}^{n} \ell(h,r_i) \tag{4}$$
$$= \sum_{k=1}^{K} \mathbb{E}_{r_k}[\ell(h,r_k)] \left(\Pr(r \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right) + \left(\sum_{k=1}^{K} \mathbb{E}_{r_k}[\ell(h,r_k)] \frac{|\mathcal{I}_k|}{n} - \frac{1}{n}\sum_{i=1}^{n} \ell(h,r_i)\right).$$

The second term in the right-hand side of (4) is further simplified by using

$$\frac{1}{n}\sum_{i=1}^{n} \ell(h,r_i) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \ell(h,r_i),$$

as

$$\sum_{k=1}^{K} \mathbb{E}_{r_k}[\ell(h,r_k)] \frac{|\mathcal{I}_k|}{n} - \frac{1}{n}\sum_{i=1}^{n} \ell(h,r_i) = \frac{1}{n}\sum_{k=1}^{K} |\mathcal{I}_k| \left(\mathbb{E}_{r_k}[\ell(h,r_k)] - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \ell(h,r_i)\right)$$

Substituting these into equation (4) yields

$$\mathbb{E}_r[\ell(h,r)] - \frac{1}{n}\sum_{i=1}^{n} \ell(h,r_i) \tag{5}$$
$$= \sum_{k=1}^{K} \mathbb{E}_{r_k}[\ell(h,r_k)] \left(\Pr(r \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right) + \frac{1}{n}\sum_{k=1}^{K} |\mathcal{I}_k| \left(\mathbb{E}_{r_k}[\ell(h,r_k)] - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \ell(h,r_i)\right)$$
$$\le B\sum_{k=1}^{K} \left|\Pr(r \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right| + \frac{1}{n}\sum_{k=1}^{K} |\mathcal{I}_k| \left(\mathbb{E}_{r_k}[\ell(h,r_k)] - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \ell(h,r_i)\right)$$

By using the Bretagnolle-Huber-Carol inequality [van der Vaart and Wellner, 1996, A6.6 Proposition], we have that for any $\delta > 0$, with probability at least $1-\delta$,

$$\sum_{k=1}^{K} \left|\Pr(r \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right| \le \sqrt{\frac{2K\ln(2/\delta)}{n}}. \tag{6}$$

Here, notice that the term of $\sum_{k=1}^{K} \left|\Pr(r \in \mathcal{C}_k) - \frac{|\mathcal{I}_k|}{n}\right|$ does not depend on $h \in \mathcal{H}$. Moreover, note that for any (f, h, M) such that $M > 0$ and $B \ge 0$ for all X, we have that $\mathbb{P}(f(X) \ge M) \ge \mathbb{P}(f(X) > M) \ge \mathbb{P}(Bf(X) + h(X) > BM + h(X))$, where the probability is with respect to the randomness of X. Thus, by combining (5) and (6), we have that for any $h \in \mathcal{H}$, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$,

$$\mathbb{E}_r[\ell(h,r)] - \frac{1}{n}\sum_{i=1}^{n} \ell(h,r_i) \le \frac{1}{n}\sum_{k=1}^{K} |\mathcal{I}_k| \left(\alpha_k(h) - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \ell(h,r_i)\right) + c\sqrt{\frac{\ln(2/\delta)}{n}}.$$

□

In particular, the first term from the previous lemma will be bounded with the following lemma:

Lemma 2. *For any* $f \in \{f^S_{\mathrm{SEM}(\tau)}, f^S_{\mathrm{base}}\}$,

$$\frac{1}{n}\sum_{k=1}^{K} |\mathcal{I}_k| \left(\alpha_k(f) - \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \ell(f,r_i)\right) \le R\sqrt{L\varphi(f)}.$$

Proof. By using the triangle inequality,

$$\frac{1}{n}\sum_{k=1}^{K}|\mathcal{I}_k|\left(\mathbb{E}_r[\ell(f,r)|r\in\mathcal{C}_k]-\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(f,r_i)\right)$$
$$\leq \frac{1}{n}\sum_{k=1}^{K}|\mathcal{I}_k|\left|\mathbb{E}_r[\ell(f,r)|r\in\mathcal{C}_k]-\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(f,r_i)\right|.$$

Furthermore, by using the triangle inequality,

$$\begin{aligned}\left|\mathbb{E}_r[\ell(f,r)|r\in\mathcal{C}_k]-\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(f,r_i)\right| &= \left|\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\mathbb{E}_r[\ell(f,r)|r\in\mathcal{C}_k]-\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(f,r_i)\right| \\ &\leq \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\left|\mathbb{E}_r[\ell(f,r)|r\in\mathcal{C}_k]-\ell(f,r_i)\right| \\ &\leq \sup_{r,r'\in\mathcal{C}_k}\left|\ell(f,r)-\ell(f,r')\right|.\end{aligned}$$

If $f=f^S_{\mathrm{SEM}(\tau)}=g^S_{\mathrm{SEM}(\tau)}\circ\sigma_\tau$, since $g^S_{\mathrm{SEM}(\tau)}\in\mathcal{G}_S$, by using the Lipschitz continuity, boundedness, and non-negativity,

$$\begin{aligned}\sup_{r,r'\in\mathcal{C}_k}\left|\ell(f,r)-\ell(f,r')\right| &= \sup_{y\in\mathcal{Y}}\sup_{z,z'\in\mathcal{Z}_k}|(l_y\circ g^S_{\mathrm{SEM}(\tau)})(\sigma_\tau(z))-(l_y\circ g^S_{\mathrm{SEM}(\tau)})(\sigma_\tau(z'))| \\ &\leq R\sup_{z,z'\in\mathcal{Z}_k}\|\sigma_\tau(z)-\sigma_\tau(z')\|_F \\ &= R\sup_{z,z'\in\mathcal{Z}_k}\sqrt{\sum_{t=1}^{L}\sum_{j=1}^{V}(\sigma_\tau(z_{t,j})-\sigma_\tau(z'_{t,j}))_2^2} \\ &\leq R\sqrt{\sum_{t=1}^{L}\sup_{i\in[V]}\sup_{q,q'\in Q_i}\|\sigma_\tau(q)-\sigma_\tau(q')\|_2^2} \\ &= R\sqrt{L\varphi(f^S_{\mathrm{SEM}(\tau)})}\end{aligned}$$

Similarly, if $f=f^S_{\mathrm{base}}=g^S_{\mathrm{base}}$, since $g^S_{\mathrm{base}}\in\mathcal{G}_S$, by using the Lipschitz continuity, boundedness, and non-negativity,

$$\begin{aligned}\sup_{r,r'\in\mathcal{C}_k}\left|\ell(f,r)-\ell(f,r')\right| &= \sup_{y\in\mathcal{Y}}\sup_{z,z'\in\mathcal{Z}_k}|(l_y\circ g^S_{\mathrm{base}})(z)-(l_y\circ g^S_{\mathrm{base}})(z')| \\ &\leq R\sup_{z,z'\in\mathcal{Z}_k}\|z-z'\|_F \\ &\leq R\sqrt{L\varphi(f^S_{\mathrm{base}})}.\end{aligned}$$

Therefore, for any $f\in\{f^S_{\mathrm{SEM}(\tau)},f^S_{\mathrm{base}}\}$,

$$\frac{1}{n}\sum_{k=1}^{K}|\mathcal{I}_k|\left(\alpha_k(f)-\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\ell(f,r_i)\right)\leq\frac{1}{n}\sum_{k=1}^{K}|\mathcal{I}_k|R\sqrt{L\varphi(f)}=R\sqrt{L\varphi(f)}.$$

□

Combining Lemma 1 and Lemma 2, we obtain the following upper bound on the gap:

Lemma 3. *For any $\delta>0$, with probability at least $1-\delta$, the following holds for any $f\in\{f^S_{\mathrm{SEM}(\tau)},f^S_{\mathrm{base}}\}$:*

$$\mathbb{E}_r[\ell(f,r)]-\frac{1}{n}\sum_{i=1}^{n}\ell(f,r_i)\leq R\sqrt{L\varphi(f)}+c\sqrt{\frac{\ln(2/\delta)}{n}}.$$

Proof. This follows directly from combining Lemma 1 and Lemma 2. $\square$

We now provide an upper bound on $\varphi(f^S_{\mathrm{SEM}(\tau)})$ in the following lemma:

Lemma 4. *For any $\tau > 0$,*

$$\varphi(f^S_{\mathrm{SEM}(\tau)}) \leq \left| \frac{1}{1+(V-1)e^{-2/\tau}} - \frac{1}{1+(V-1)e^{-\Delta/\tau}} \right|^2 + (V-1) \left| \frac{1}{1+e^{\Delta/\tau}(1+(V-2)e^{-2/\tau})} - \frac{1}{1+e^{2/\tau}(1+(V-2)e^{-\Delta/\tau})} \right|^2.$$

Proof. Recall the definition:

$$\varphi(f^S_{\mathrm{SEM}(\tau)}) = \sup_{i\in[V]} \sup_{q,q'\in Q_i} \|\sigma_\tau(q) - \sigma_\tau(q')\|_2^2.$$

where

$$\sigma_\tau(q)_j = \frac{e^{q_j/\tau}}{\sum_{t=1}^V e^{q_t/\tau}},$$

for $j = 1, \ldots, V$. By the symmetry and independence over $i \in [V]$ inside of the first supremum, we have

$$\varphi(f^S_{\mathrm{SEM}(\tau)}) = \sup_{q,q'\in Q_1} \|\sigma_\tau(q) - \sigma_\tau(q')\|_2^2.$$

For any $q, q' \in Q_1$ and $i \in \{2, \ldots, V\}$ (with $q = (q_1, \ldots, q_V)$ and $q' = (q'_1, \ldots, q'_V)$), there exists $\delta_i, \delta'_i > 0$ such that

$$q_i = q_1 - \delta_i$$

and

$$q'_i = q'_1 - \delta'_i.$$

Here, since $z_{ik} - \Delta \geq z_{ij}$ from the assumption, we have that for all $i \in \{2, \ldots, V\}$,

$$\delta_i, \delta'_i \geq \Delta > 0.$$

Thus, we can rewrite

$$\begin{aligned}\sum_{t=1}^V e^{q_t/\tau} &= e^{q_1/\tau} + \sum_{i=2}^V e^{(q_1-\delta_i)/\tau} \\ &= e^{q_1/\tau} + e^{q_1/\tau}\sum_{i=2}^V e^{-\delta_i/\tau} \\ &= e^{q_1/\tau}\left(1 + \sum_{i=2}^V e^{-\delta_i/\tau}\right)\end{aligned}$$

Similarly,

$$\sum_{t=1}^V e^{q'_t/\tau} = e^{q'_1/\tau}\left(1 + \sum_{i=2}^V e^{-\delta'_i/\tau}\right).$$

Using these,

$$\sigma_\tau(q)_1 = \frac{e^{q_1/\tau}}{\sum_{t=1}^V e^{q_t/\tau}} = \frac{e^{q_1/\tau}}{e^{q_1/\tau}\left(1+\sum_{i=2}^V e^{-\delta_i/\tau}\right)} = \frac{1}{1+\sum_{i=2}^V e^{-\delta_i/\tau}}$$

and for all $j \in \{2, \ldots, V\}$,

$$\sigma_\tau(q)_j = \frac{e^{q_j/\tau}}{\sum_{t=1}^V e^{q_t/\tau}}$$
$$= \frac{e^{(q_1-\delta_j)/\tau}}{e^{q_1/\tau}\left(1 + \sum_{i=2}^V e^{-\delta_i/\tau}\right)}$$
$$= \frac{e^{-\delta_j/\tau}}{1 + \sum_{i=2}^V e^{-\delta_i/\tau}}$$
$$= \frac{1}{1 + e^{\delta_j/\tau} + \sum_{i\in I_j}^V e^{(\delta_j-\delta_i)/\tau}}$$

where $I_j := \{2, \ldots, V\} \setminus \{j\}$. Similarly,

$$\sigma_\tau(q')_1 = \frac{1}{1 + \sum_{i=2}^V e^{-\delta'_i/\tau}},$$

and for all $j \in \{2, \ldots, V\}$,

$$\sigma_\tau(q')_j = \frac{1}{1 + e^{\delta'_j/\tau} + \sum_{i\in I_j}^V e^{(\delta'_j-\delta'_i)/\tau}}.$$

Using these, for any $q, q' \in Q_1$,

$$|\sigma_\tau(q)_1 - \sigma_\tau(q')_1| = \left|\frac{1}{1 + \sum_{i=2}^V e^{-\delta_i/\tau}} - \frac{1}{1 + \sum_{i=2}^V e^{-\delta'_i/\tau}}\right|$$
$$\leq \left|\frac{1}{1 + \sum_{i=2}^V e^{-2/\tau}} - \frac{1}{1 + \sum_{i=2}^V e^{-\Delta/\tau}}\right|$$
$$= \left|\frac{1}{1 + (V-1)e^{-2/\tau}} - \frac{1}{1 + (V-1)e^{-\Delta/\tau}}\right|,$$

and for all $j \in \{2, \ldots, V\}$,

$$|\sigma_\tau(q)_j - \sigma_\tau(q')_j| = \left|\frac{1}{1 + e^{\delta_j/\tau} + \sum_{i\in I_j}^V e^{(\delta_j-\delta_i)/\tau}} - \frac{1}{1 + e^{\delta'_j/\tau} + \sum_{i\in I_j}^V e^{(\delta'_j-\delta'_i)/\tau}}\right|$$
$$\leq \left|\frac{1}{1 + e^{\Delta/\tau} + \sum_{i\in I_j}^V e^{(\Delta-2)/\tau}} - \frac{1}{1 + e^{2/\tau} + \sum_{i\in I_j}^V e^{(2-\Delta)/\tau}}\right|$$
$$= \left|\frac{1}{1 + e^{\Delta/\tau} + (V-2)e^{(\Delta-2)/\tau}} - \frac{1}{1 + e^{2/\tau} + (V-2)e^{(2-\Delta)/\tau}}\right|$$
$$= \left|\frac{1}{1 + e^{\Delta/\tau}(1 + (V-2)e^{-2/\tau})} - \frac{1}{1 + e^{2/\tau}(1 + (V-2)e^{-\Delta/\tau})}\right|.$$

By combining these,

$$\sup_{q,q'\in Q_1} \|\sigma_\tau(q) - \sigma_\tau(q')\|_2^2$$
$$= \sup_{q,q'\in Q_1} \sum_{j=1}^V |\sigma_\tau(q)_j - \sigma_\tau(q')_j|^2$$
$$\leq \left|\frac{1}{1 + (V-1)e^{-2/\tau}} - \frac{1}{1 + (V-1)e^{-\Delta/\tau}}\right|^2$$
$$\quad + (V-1)\left|\frac{1}{1 + e^{\Delta/\tau}(1 + (V-2)e^{-2/\tau})} - \frac{1}{1 + e^{2/\tau}(1 + (V-2)e^{-\Delta/\tau})}\right|^2.$$

$\square$

Using the previous lemma, we will conclude the asymptotic behavior of $\varphi(f^S_{\mathrm{SEM}(\tau)})$ in the following lemma:

Lemma 5. *It holds that*

$$\varphi(f^S_{\mathrm{SEM}(\tau)}) \to 0 \text{ as } \tau \to 0.$$

Proof. Using Lemma 4,

$$\begin{aligned}\lim_{\tau\to 0}\varphi(f^S_{\mathrm{SEM}(\tau)}) \le \lim_{\tau\to 0}&\left|\frac{1}{1+(V-1)e^{-2/\tau}} - \frac{1}{1+(V-1)e^{-\Delta/\tau}}\right|^2 \\ &+ n(V-1)\lim_{\tau\to 0}\left|\frac{1}{1+e^{\Delta/\tau}(1+(V-2)e^{-2/\tau})} - \frac{1}{1+e^{2/\tau}(1+(V-2)e^{-\Delta/\tau})}\right|^2.\end{aligned}$$

Moreover,

$$\lim_{\tau\to 0}\left|\frac{1}{1+(V-1)e^{-2/\tau}} - \frac{1}{1+(V-1)e^{-\Delta/\tau}}\right|^2 = \left|\frac{1}{1}-\frac{1}{1}\right|^2 = 0,$$

and

$$\lim_{\tau\to 0}\left|\frac{1}{1+e^{\Delta/\tau}(1+(V-2)e^{-2/\tau})} - \frac{1}{1+e^{2/\tau}(1+(V-2)e^{-\Delta/\tau})}\right|^2 = |0-0|^2 = 0.$$

Therefore,

$$\lim_{\tau\to 0}\varphi(f^S_{\mathrm{SEM}(\tau)}) \le 0.$$

Since $\varphi(f^S_{\mathrm{SEM}(\tau)}) \ge 0$, this implies the statement of this lemma. $\square$

As we have analyzed $\varphi(f^S_{\mathrm{SEM}(\tau)})$ in the previous two lemmas, we are now ready to compare $\varphi(f^S_{\mathrm{SEM}(\tau)})$ and $\varphi(f^S_{\mathrm{base}})$, which is done in the following lemma:

Lemma 6. *For any $\tau > 0$,*

$$\varphi(f^S_{\mathrm{SEM}(\tau)}) - \varphi(f^S_{\mathrm{base}}) \le \frac{3}{4}(1-V) < 0.$$

Proof. From Lemma 4, for any $\tau > 0$,

$$\begin{aligned}\varphi(f^S_{\mathrm{SEM}(\tau)}) &\le \left|\frac{1}{1+(V-1)e^{-2/\tau}} - \frac{1}{1+(V-1)e^{-\Delta/\tau}}\right|^2 \\ &\quad + n(V-1)\left|\frac{1}{1+e^{\Delta/\tau}(1+(V-2)e^{-2/\tau})} - \frac{1}{1+e^{2/\tau}(1+(V-2)e^{-\Delta/\tau})}\right|^2 \\ &\le \left|\frac{1}{1+(V-1)e^{-2/\tau}} - \frac{1}{1+(V-1)}\right|^2 \\ &\quad + (V-1)\left|\frac{1}{1+(1+(V-2)e^{-2/\tau})} - \frac{1}{1+e^{2/\tau}(1+(V-2))}\right|^2 \\ &= \left|\frac{1}{1+(V-1)e^{-2/\tau}} - \frac{1}{V}\right|^2 + (V-1)\left|\frac{1}{2+(V-2)e^{-2/\tau}} - \frac{1}{1+e^{2/\tau}(V-1)}\right|^2 \\ &\le \left|\frac{1}{1}-\frac{1}{V}\right|^2 + (V-1)\left|\frac{1}{2}-0\right|^2 \\ &= \left(\frac{1}{1}-\frac{1}{V}\right)^2 + (V-1)\frac{1}{4}.\end{aligned}$$

Recall the definition of

$$\varphi(f^S_{\text{base}}) = \sup_{i \in [V]} \sup_{q,q' \in Q_i} \|q - q'\|_2^2.$$

By choosing an element in the set over which the supremum is taken, for any $\delta \geq \Delta > 0$,

$$\varphi(f^S_{\text{base}}) \geq \sup_{q,q' \in Q_1} \|q - q'\|_2^2 \geq \|\hat{q} - \hat{q}'\|_2^2 = \sum_{j=1}^{V} (\hat{q}_j - \hat{q}'_j)_2^2 = (2-\delta)^2 V,$$

where $\hat{q}_1 = 1$, $\hat{q}_j = 1 - \delta$ for $j \in \{2, \ldots, V\}$, $\hat{q}'_1 = \delta - 1$, and $\hat{q}'_j = -1$ for $j \in \{2, \ldots, V\}$.

By combining those, for for any $\tau > 0$ and $\delta \geq \Delta > 0$,

$$\begin{aligned}
\varphi(f^S_{\text{SEM}(\tau)}) - \varphi(f^S_{\text{base}}) &\leq \left(\frac{1}{1} - \frac{1}{V}\right)^2 + (V-1)\frac{1}{4} - (2-\delta)^2 V \\
&\leq 1 + \frac{1}{4}V - \frac{1}{4} - (2-\delta)^2 V \\
&= \frac{3}{4} + \frac{1}{4}V - (2-\delta)^2 V \\
&= \frac{3}{4} - V\left((2-\delta)^2 - \frac{1}{4}\right) \\
&\leq \frac{3}{4} - V\left(1 - \frac{1}{4}\right) \\
&= \frac{3}{4}(1 - V)
\end{aligned}$$

□

We combine the lemmas above to prove Theorem 1, which is restated below with its proof:

Theorem 1. *Let $V \geq 2$. For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for any $f_S \in \{f^S_{\text{SEM}(\tau)}, f^S_{\text{base}}\}$:*

$$\mathbb{E}_{z,y}[l(f_S(z), y)] \leq \frac{1}{n}\sum_{i=1}^{n} l(f_S(z^{(i)}), y^{(i)}) + R\sqrt{L\varphi(\sigma_{f_S})} + c\sqrt{\frac{\ln(2/\delta)}{n}}.$$

Moreover,

$$\varphi(\sigma_{f^S_{\text{SEM}(\tau)}}) \to 0 \quad \textit{as } \tau \to 0 \quad \textit{and} \quad \varphi(\sigma_{f^S_{\text{SEM}(\tau)}}) - \varphi(\sigma_{f^S_{\text{base}}}) \leq \frac{3}{4}(1 - V) < 0 \quad \forall \tau > 0.$$

Proof. The first statement directly follows from Lemma 3. The second statement is proven by Lemma 5 and Lemma 6. □

C Additional experiments on CIFAR-100

Increasing L increases the performance of SEM. We find that increasing L, the number of simplex of SEM even beyond the over-complete regime increases the downstream accuracy. However, this increased performance is not observed when we abstain from using the softmax normalization of SEM. In Figure 3, using a ResNet-50 encoder, we compare BYOL + SEM, with an identical model without the Softmax normalization which we call BYOL + Embed. As, this is a control experiment, the extracted representation of BYOL + Embed is the embedder's output z_θ. We fix $V = 13$ and scale $L \in [10, 10000]$ to get a range of representation sizes. We observe that BYOL + Embed accuracy's is decrease from probing z_θ instead of e_θ showing no gain from larger representations.

Memory and computational efficiency of SEM. We present the memory requirement and the computational efficiency of SEM in Table 5. The allocated memory represent the VRAM allocated by PyTorch during pre-training with a Batch Size of 256. As expected, SEM necessitates much more memory, which can become a practical issue. Fortunately, sparsifying the matrix multiplication as discussed in Section A.2, allows to considerably reduce the memory requirements while inducing a small reduction in performance. In term of computational efficiency, we note that the cost of using SEM is small in comparison to the total cost of the pre-training and becomes marginal as we scale up the encoder.

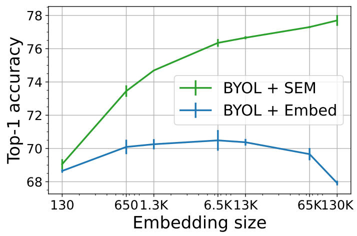


Figure 3: Effect of the Softmax when scaling up L of SEM. Using a RN-50 encoder.

	vRAM (GiB)	FLOPs	Acc.
Resnet-18:			
BYOL	4.0	7.20*e*8	70.7
BYOL+SEM	13.1	1.01*e*9	73.9
BYOL+SEM/4	6.2	7.83*e*8	73.3
Resnet-50:			
BYOL	11.1	1.65*e*9	74.3
BYOL+SEM	21.9	2.04*e*9	77.4
BYOL+SEM/4	13.5	1.74*e*9	76.1

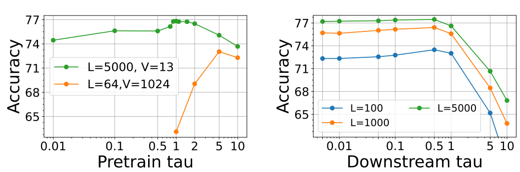
Table 5: Allocated memory, computation efficiency (calculated in FLOPs/sample) and accuracy of BYOL with SEM.

	Acc.
BYOL	70.5
BYOL+REINFORCE	5.60
BYOL+Gumbel S.T.	45.9
BYOL+V.Q.	56.7
BYOL+SEM($\tau_{\mathrm{d}} = 0$)	73.2
BYOL+SEM($\tau_{\mathrm{d}} = 0.1$)	**73.9**

Table 6: Comparing SEM with REINFORCE, Gubel Straight-Through and Vector Quantization (V.Q.). Using a RN-18 encoder.

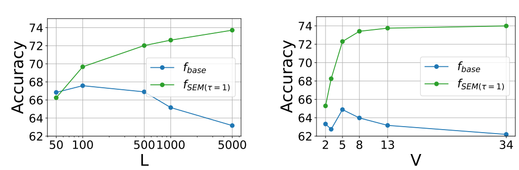
Comparison of SEM with *hard* discretization approaches. Several other methods can be used to induce a sparse representation during pre-training and downstream classification. For example, we may sample L discrete one-hot codes each with V values using REINFORCE [Williams, 1992] or Gumbel Straight-Through estimation [Jang et al., 2017] to back-propagate the gradient through the encoder. We could also use Vector Quantization (VQ) [Oord et al., 2018] as discussed in Liu et al. [2021] and consider L codebooks with V values each wherein the values are vectors in $\mathbb{R}^d$. However, when using these approaches instead of SEM, we see a considerable decrease in performance in comparison to the baseline as demonstrated in Table 6. In this table, we reproduce the same setup as SEM but we replace the Softmax with hard discretization baselines methods. For discretization with REINFORCE and Gumbel Straight-Through estimation, we use the same setup as SEM with $L = 5000$ and $V = 13$, that is 5000 one-hot vectors of 13 dimensions and $\tau = 1$ for Gumbel Straight-Through. For VQ, we found that $L = 512$ and $V = 128$ led to the best performance. That is, we have 512 codebooks, each with 128 possible values that are represented by vectors in $\mathbb{R}^{32}$. We also present SEM with $\tau_{\mathrm{DS}} = 0$, which correspond to using the discretized representation for downstream classification, demonstrating that SEM with pre-training can be used to learn meaningful discrete codes for downstream task and yields better performance than the baselines, leading us to believe that SEM could be beneficial in setup where discretization is needed and pre-training is possible.

C.1 Analyzing the parameters of SEM



(a) (b)

Figure 4: Effect of τ_p and τ_d on a RN-50.



(a) (b)

Figure 5: Comparing f_{SEM} and f_{base} on a RN-18.

We present two figures in this section to better understand the effect of the parameters of SEM on the downstream accuracy. In Figure 4, we evaluate the effect of changing τ_p and τ_d on the downstream accuracy. In Figure 5, we evaluate the effect of L and V on the downstream accuracy and also contrast f_{base} and $f_{\mathrm{SEM}}(\tau = 1)$, allowing us to confirm two predictions made in Theorem 1. The expected generalization improvement from SEM increases as we increase L and as we increase V. We now discuss the effect of each of SEM's parameter on the resulting downstream classification.

Increasing V yields steep performance increase for small V but quickly plateau. In Figure 5b, we observe a steep increase of the accuracy for $V < 13$ followed by a plateau for $V > 13$. In Figure 4a, we observe that the optimal accuracy obtained for $V = 1024$ and $L = 64$ is similar to the one obtained for $L = 50$ (Embedding size=650) in Figure 3.

Superclass	Classes
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food containers	bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household furniture	bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

Table 7: Set of classes for each superclass on CIFAR-100.

Increasing L yields monotonical improvement. In the regime that we can test it, larger L seems to translate to better accuracy as observed in Figure 3 and Figure 5a.

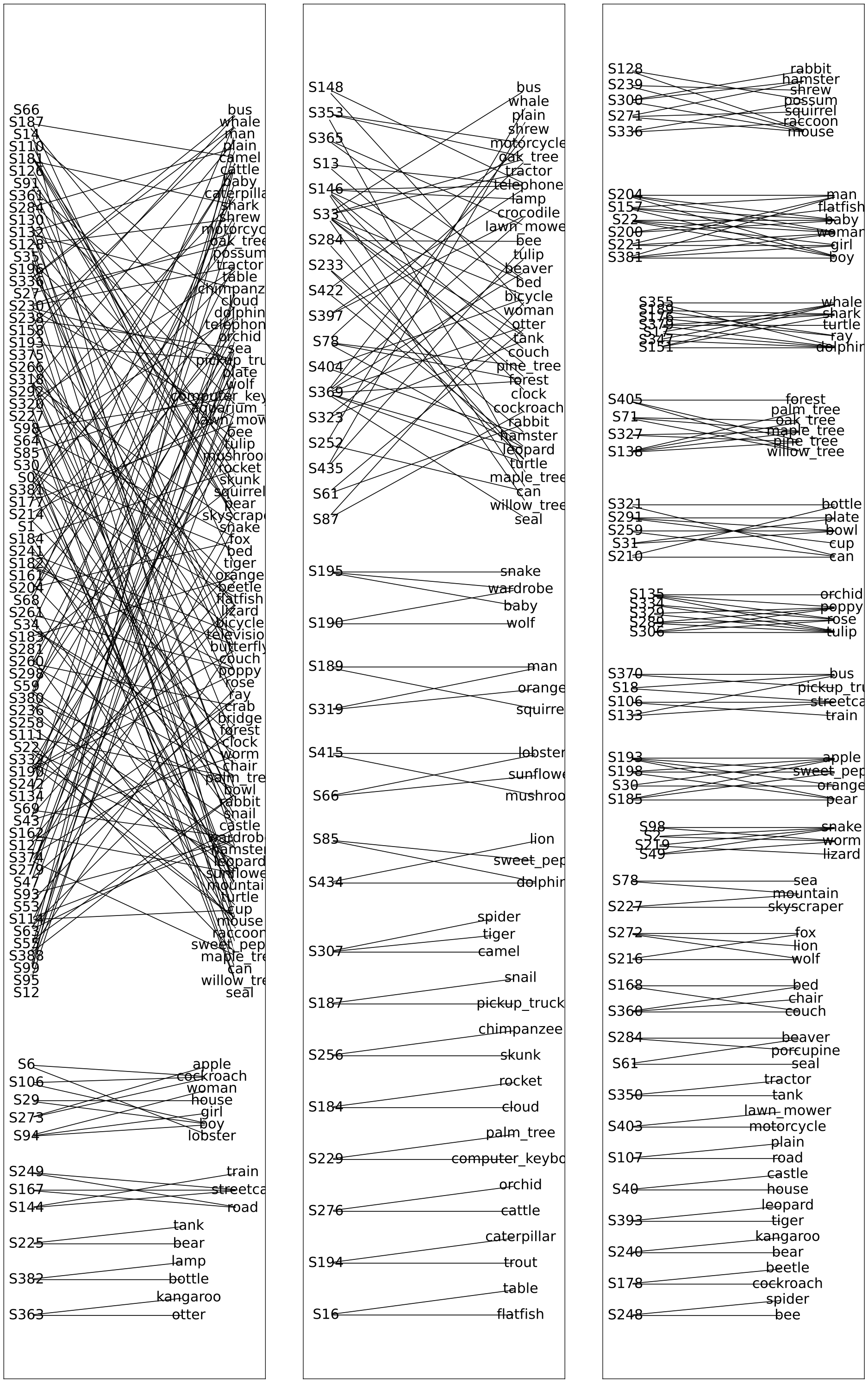
The optimal τ_p depends on V. As previously noted in the context of Attention [Vaswani et al., 2017; Wang et al., 2021a], the optimal attention's temperature is proportional to attention's vector size. This is also observed in SEM. As presented in Figure 4a, the optimal τ_p for larger V is higher.

Models with larger L are more robust to smaller τ_d. In Figure 4, we observe that SSL models are more robust to smaller τ_d as L increase. As L is larger, we speculate that the information can be scattered across the simplices, allowing to reduce the expressivity of each vector with minimal impact on the downstream accuracy.

D CIFAR100 superclass

The 100 classes of CIFAR-100 [Krizhevsky, 2009] are grouped into 20 superclasses. The list of superclass for each class in Table 7

E Additional CIFAR-100 coherence graphs



(a) BYOL baseline

(b) BYOL baseline with a large representation

(c) BYOL + SEM

Figure 6: Comparison of the full semantic coherence graph $\mathcal{W}_5$ between BYOL and BYOL + SEM.