Evaluating and Improving Robustness of Self-Supervised Representations to Spurious Correlations

Anonymous Authors[1]

Abstract

Recent empirical studies have found inductive biases in supervised learning toward simple features that may be spuriously correlated with the label, resulting in suboptimal performance on certain subgroups. In this work, we explore whether recent Self-Supervised Learning (SSL) methods would produce representations which exhibit similar behavior. First, we show that classical approaches in combating spurious correlations, such as re-sampling the dataset, do not necessarily lead to invariant representations during SSL. Second, we discover that spurious information is represented disproportionately heavily in the later layers of the encoder. Motivated by these findings, we propose a method to remove spurious information from SSL representations during pretraining, by pruning or re-initializing later layers of the encoder. We find that our method produces representations which outperform the baseline on 5 datasets, without the need for group or label information during SSL.

1. Introduction

Many real-world predictive tasks contain spurious correlations – features that are correlated with the label only for certain subsets of the data (Shah et al., 2020; McCoy et al., 2019; Gururangan et al., 2018). In such cases, Empirical Risk Minimization (ERM) models make predictions based on correlations in the data which generalizes on average in the training set but results in systematically poor performance for minority subgroups (Hashimoto et al., 2018). Such correlations may not exist in future test distributions, leading to performance drops in test distributions. For example, models trained to detect pneumonia (Zech et al., 2018) or COVID-19 (DeGrave et al., 2021) from chest X-rays across hospitals used the hospital ID as a shortcut instead of focusing on the invariant pulmonary characteristics.

There have been many methods proposed to tackle spurious correlations for supervised learning. Some methods require group information during training (Sagawa et al., 2020a), while others do not (Liu et al., 2021a; Zhang et al., 2022). However, all methods require group information for model selection.

Recently, there has been a rise in the popularity of Self-Supervised Learning (SSL) methods, which seek to learn data representations from large, *unlabeled* datasets (Chen et al., 2020a; He et al., 2019; Grill et al., 2020; Chen & He, 2020; Caron et al., 2020; Zbontar et al., 2021). In this work, we ask the following questions: (1) Do SSL methods also have a bias towards learning representations which result in spurious correlations when a linear ERM model is trained on representations for downstream tasks? (2) If so, can we suppress such features from learnt representations during SSL, without access to any group or label information?

We make the following contributions:

- We show that simple techniques for avoiding spurious correlations during supervised learning, such as re-weighting (Idrissi et al., 2021) or re-sampling (Sagawa et al., 2020b) of the training set with group information, does not improve representations learnt with SSL.
- We show that group information is correlated with features learned in the final layers of the network, and hypothesize that removing such information from the final layers can be beneficial.
- We propose a method to eliminate spurious feature information from representations learnt during SSL while maintaining discriminative ability for downstream predictive tasks, without access to group or label information.

2. Related Work

Spurious Correlations Spurious correlations arise in supervised learning models in a variety of domains, from medical imaging (Zech et al., 2018; DeGrave et al., 2021)

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

Dataset	SSL Training	kNN (Target)		LR (Target)		LR (Group)
		Average	Worst Group	Average	Worst Group	Average
waterbirds	Normal	50.5%	46.7%	53.2%	51.8%	86.2%
	Downsample	51.7%	47.0%	55.5%	50.5%	85.3%
	Upsample	66.3%	9.5%	62.3%	46.7%	81.5%
metashift	Normal	52.5%	22.8%	65.8%	13.6%	62.4%
	Downsample	45.1%	0.0%	59.8%	0.0%	57.4%
	Upsample	56.5%	32.9%	65.8%	24.2%	65.1%
spurcifar10	Normal	57.1%	18.8%	60.6%	38.5%	49.0%
	Downsample	36.0%	19.6%	48.8%	28.8%	48.3%
	Upsample	43.8%	20.7%	53.6%	9.3%	70.7%
cmnist	Normal	89.5%	40.7%	86.2%	38.9%	84.8%
	Downsample	86.2%	23.7%	80.0%	35.2%	83.2%
	Upsample	86.9%	39.4%	80.0%	43.7%	79.6%

Table 1: Accuracy of ResNet-18 encoders trained using SSL on four datasets, evaluated on a balanced test set. We vary the SSL training set used (Normal: original dataset, Downsample: downsampling majority groups, Upsample: upsampling minority groups). Re-sampling does not necessarily improve worst-group accuracy in the downstream task.

to natural language processing (Tu et al., 2020; Wang & Culotta, 2020). A variety of approaches have been proposed to learn classifiers which do not make use of spurious information. Methods like GroupDRO (Sagawa et al., 2020a) and DFR (Kirichenko et al., 2022) require group information during training, while methods like JTT (Liu et al., 2021a), LfF (Nam et al., 2020), CVaR DRO (Duchi et al., 2019), CnC (Zhang et al., 2022) do not. However, all methods require group information for model selection.

Self-supervised Representation Learning Self-supervised learning methods learn representations from large-scale unlabeled datasets where annotations are scarce and time-consuming. In the image setting, the pretext task is typically to maximize similarity between an image and its augmented view, or between two augmented views of the same image (Jing & Tian, 2020). This can be done in a contrastive fashion using the InfoNCE loss (Oord et al., 2018), as in SimCLR (Chen et al., 2020a) and MoCo (Chen et al., 2020b), or without the need for negative samples at all, as in BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021).

Dataset Imbalance and Shortcut Learning Shortcut learning has been identified in contrastive learning (Robinson et al., 2021), and addressed by adversarially modifying encoded features. Even though self-supervised models have been found to be more robust to dataset imbalance (Liu et al., 2021b), Jiang et al. (2021b;a); Liu et al. (2021b) further address the gap. Other works in addressing group robustness or fairness in SSL, however, require group information or labels (Tsai et al., 2020; Song et al., 2019). Similar to our work, improving generalization of the model through forgetting events via final-layer re-initialization, has been studied in the supervised setting (Zhou et al., 2022).

For a more comprehensive summary of the background and related work, see Appendix A.

3. Motivation

3.1. SSL Suffers From Spurious Correlations

We empirically evaluate learning of the invariant and spurious features on self-supervised representations. We train ResNet-18 encoders using SimSiam (Chen & He, 2020) on 4 datasets containing spurious correlations (See Appendix B for dataset descriptions), and evaluate the learned representations using a balanced dataset which we create by subsampling majority groups (Sagawa et al., 2020b; Idrissi et al., 2021). We use the balanced dataset to separately train two linear classifiers to distinguish labels and spurious attribute. From Table 1, we find that across all datasets, the worst-group accuracy and average accuracy for learning the spurious feature is higher. This motivates methods to remove spurious information from representations during SSL.

3.2. Re-sampling During SSL is Not Useful

In the next step, we examine whether classical approaches for combating spurious correlations, such as re-sampling training examples, are effective in removing spurious information during SSL. Assuming that group and label information are available, we train SimSiam on datasets re-sampled

using the following strategies: (i) downsampling examples in majority groups to have the same number of examples in all groups, (ii) upsampling minority examples to have the same number of examples in all groups. From Table 1, we observe that re-sampling during SSL does not improve worst-group accuracy, and the model still has a better performance in learning the spurious feature.

3.3. Layer-Wise Feature Representations

Given that conventional methods for combating spurious correlations during supervised learning do not work during SSL, we seek to come up with a method that is tailored to the SSL setting.

One crucial observation from prior work is that during supervised learning, overparameterized models have an inductive bias towards learning the spurious feature and "memorizing" the minority examples (Sagawa et al., 2020b). It has been also observed in the supervised setting that memorization predominately occurs in the deeper layers (Zhang et al., 2019; Stephenson et al., 2021) and that the representations in earlier layers of the network correctly classify the easier examples while the final layers memorize the difficult examples (Baldock et al., 2021).

We suspect that a similar phenomenon happens during SSL, as the spurious feature is easier to infer, and almost sufficient for the instance discrimination task. To verify this hypothesis, we evaluate the mutual information between feature representations across the layers of the trained encoder with (1) the labels $I(Z;Y)$, and (2) the spurious attribute $I(Z;G)$.

From Figure 1, we see that spurious features are disproportionately represented in later layers of the network, while invariant features are represented throughout the network. This confirms our hypothesis that later layers contain more spurious information during SSL, and motivates our proposed method.

4. Method

4.1. Implicit Representation Augmentations

In the previous sections, we showed how overparameterized representation encoders can fail in cases where there are certain minority groups within the training samples. We further empirically showed that the final layers of the model tend to be more predictive of the spurious feature. Moreover, recent findings on when these models fail to generalize to minority groups (Hooker et al., 2020), suggest that lower-magnitude weights tend to be responsible for classifying these examples. Motivated by these observation, we define a set of transformations; including re-initialization of the final layers, and pruning the final layers, on the encoder network,

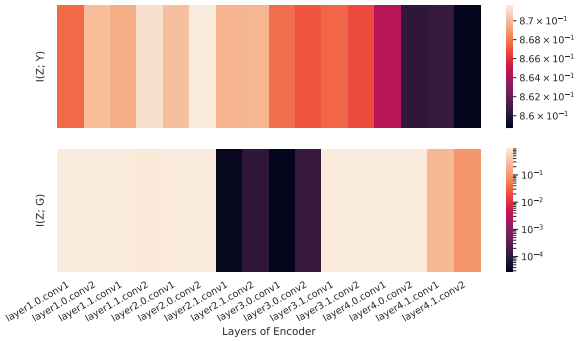

Figure 1: Comparison of Mutual Information between features and labels (top) vs. spurious attribute and labels (bottom). The first few layers are in charge of low-level features and statistics of the data, in the intermediate layers in the bottom figure, $I(Z;G)$ decreases, and in the final layers the representations rely on the spurious feature for the instance discrimination task in SSL.

which the model should be consistent in.

More formally, given two randomly augmented views x_1 and x_2 from an image x, and a feature encoder f, at each step, given a transformed model $\tilde{f}$ that heavily effects the minority examples, or *removes* the undesired information, we want its outputs to be close to the original encoded examples in the representation space. Suppose that we have such transformation. During the SSL training, one example x_1 is passed through the normal encoder $v_1 = f(x_1)$, and the other example x_2 is passed through the transformed encoder $\tilde{v}_2 = \tilde{f}(x_2)$. Encoded feature $\tilde{v}_2$ is now a positive example that should be close to v_1 in the representation space. In other words, such model transformations can be considered as curated view generating operations for the minority groups.

The transformations that we use based on insights from the previous section, for an encoder f with intermediate layers $\{f_1, f_2, \ldots, f_N = f\}$ are:

- Threshold Pruning: Magnitude pruning $a\%$ of the weights in all layers $l < L$ deeper than L ($\text{SSL}_{\text{Thres, L, a}}$)
- Re-initialization of the final-layers: Re-initializing the weights in layers $l < L$ deeper than l ($\text{SSL}_{\text{Reinit, L}}$)

We hypothesize that these transformations force the encoder to represent the features in the earlier layers of the network. For instance, when re-initializing, examples should have a consistent representation for different possible values of the final layers. Thus, the undesirable information associated with difficult or minority examples is removed which leads representing more complex features.

Dataset	Method	kNN (Target)		LR (Target)		LR (Group)	Other Metrics		
		Worst Group	Average	Worst Group	Average	Average	I(Z; Y)	I(Z; G)	L_{align}
cmnist	SSL_{Base}	44.4%	85.6%	39.0%	89.8%	84.5%	8.59E-01	8.82E-05	1.129
	$SSL_{Reinit, 15}$	33.90%	86.19%	42.37%	78.89%	79.72%	8.42E-01	8.89E-01	0.511
	$SSL_{Thres, 0.7, 19}$	37.04%	88.49%	50.00%	81.62%	82.07%	8.60E-01	9.03E-01	1.21
	$SSL_{Thres, 0.9, 18}$	32.20%	89.04%	44.07%	84.32%	81.55%	8.54E-01	9.02E-01	1.12
	$SSL_{Thres, 0.99, 19}$	35.19%	88.31%	44.07%	79.64%	82.00%	8.51E-01	8.73E-01	0.947
metashift	SSL_{Base}	18.6%	65.5%	31.0%	57.5%	62.9%	8.41E-02	8.43E-02	0.925
	$SSL_{Reinit, 18}$	21.38%	61.23%	18.64%	62.95%	66.52%	1.54E-03	1.86E-01	0.940
	$SSL_{Thres, 0.7, 17}$	27.59%	60.09%	18.64%	63.09%	67.24%	1.31E-01	1.89E-01	0.958
	$SSL_{Thres, 0.8, 17}$	24.83%	58.66%	20.34%	64.95%	64.66%	7.37E-02	1.58E-01	0.955
	$SSL_{Thres, 0.9, 18}$	25.52%	59.23%	18.64%	64.66%	65.52%	3.85E-02	1.46E-01	0.947
spurcifar10	SSL_{Base}	27.1%	54.7%	11.1%	44.0%	61.7%	1.51E+00	4.68E-03	0.750
	$SSL_{Thres, 0.3, 18}$	31.25%	61.60%	45.83%	66.86%	48.91%	1.98E+00	1.10E-05	1.08
	$SSL_{Thres, 0.5, 18}$	30.77%	58.44%	43.75%	66.23%	49.37%	2.06E+00	1.98E-09	1.05
	$SSL_{Thres, 0.7, 19}$	35.42%	58.06%	41.67%	65.80%	49.79%	2.03E+00	2.30E-02	1.04
	$SSL_{Thres, 0.9, 17}$	25.00%	50.98%	39.58%	58.59%	48.66%	1.65E+00	5.56E-05	0.630
waterbirds	SSL_{Base}	43.4%	50.2%	46.2%	49.5%	84.4%	8.23E-03	5.41E-01	0.706
	$SSL_{Thres, 0.3, 17}$	48.29%	50.59%	52.34%	55.59%	86.54%	5.03E-02	5.92E-01	0.858
	$SSL_{Thres, 0.7, 17}$	43.77%	54.99%	51.56%	62.10%	81.76%	7.86E-02	5.72E-01	0.873
	$SSL_{Thres, 0.8, 18}$	35.83%	60.23%	50.78%	60.10%	84.10%	1.01E-01	4.60E-01	0.802
	$SSL_{Thres, 0.9, 17}$	47.54%	50.88%	52.18%	57.82%	89.80%	9.63E-02	0.595E-01	0.702

Table 2: Top model transformations for each dataset; The learned representations in each case, we freeze the representations to evaluate the representations with: (i) average and worst-group of a 5-NN classifier (ii) Average and worst-group of a linear classifier (iii) Average accuracy of a linear classifier trained to infer the spurious feature (iv) Mutual information between representations and classes (v) Mutual information between representations and spurious feature (vi) Alignment loss (Zhang et al., 2022) which is indicator of how close examples within the same class are in the representation space. In many cases, the worst-group accuracy of linear classifier is improved; in other cases kNN accuracy has improvements;

5. Results

We train SimSiam models with ResNet-18 backbones on four spuriously correlated datasets (See Appendix B for dataset descriptions). For simplicity, We keep the pruning percentage and re-initialization or pruning threshold fixed throughout training. At each step t, given these hyperparameters, the fixed transformation is applied to encoder f_t to generate the transformed model $\tilde{f}_t$. The model transformations are applied to the branch of SimSiam that a gradient-stop operation is applied to later. We use group information in the validation set and measure worst-group accuracy in order to choose layer threshold and pruning percentage hyperparameters (In Appendix D we show that even an uncurated set of model transformations enhance worst-group performance)

Then, we evaluate the learned representations using a balanced dataset which we create by subsampling majority groups, to avoid the statistical and geometrical skews (Nagarajan et al., 2020) when training the linear classifier.

Table 2 shows that the worst-group accuracy of the linear classifier is improved by up to more than 30% on `spurcifar10`. However, it does not improve the worst-group accuracy in `metashift`. The cats vs dogs task in the metashift datset is originally created by gathering different subsets of pet images that appeared in an image with an object associated with indoor or outdoor. We suspect that the spurious feature in this case, is not significantly easier to infer, which would lead to poor performance i any approach that would assume that the spurious feature is easier to infer.

6. Discussion

We studied the performance of SSL models when trained on spuriously correlated data, and showed that the invariant features is not very well distinguished in many cases. We further examined whether re-sampling approaches can help mitigate this problem. We show that without having access to group or label information, the worst-group downstream performance can be improved.

References

Adel, T., Valera, I., Ghahramani, Z., and Weller, A. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2412–2420, 2019.

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., and Rus, D. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289–295, 2019.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. 2019. URL http://arxiv.org/abs/1902.09229. cite arxiv:1902.09229Comment: 19 pages, 5 figures.

Baldock, R., Maennel, H., and Neyshabur, B. Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems*, 34, 2021.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Chen, R. T., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

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

Chen, X. and He, K. Exploring Simple Siamese Representation Learning. *arXiv e-prints*, art. arXiv:2011.10566, November 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

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

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Duchi, J. C., Hashimoto, T., and Namkoong, H. Distributionally robust losses against mixture covariate shifts. *Under review*, 2, 2019.

Edwards, H. and Storkey, A. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and Meent, J.-W. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2525–2534. PMLR, 2019.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19: 513–520, 2006.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z. D., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv e-prints*, art. arXiv:2006.07733, June 2020.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. *arXiv:1911.05722*, 2019.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. *arXiv preprint arXiv:2110.14503*, 2021.

Jiang, Z., Chen, T., Chen, T., and Wang, Z. Improving contrastive learning on imbalanced data via open-world sampling. *Advances in Neural Information Processing Systems*, 34, 2021a.

Jiang, Z., Chen, T., Mortazavi, B., and Wang, Z. Self-damaging contrastive learning. *arXiv preprint arXiv:2106.02990*, 2021b.

Jing, L. and Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 4037–4058, 2020.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*, 2022.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021a.

Liu, H., HaoChen, J. Z., Gaidon, A., and Ma, T. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021b.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

McCoy, R. T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., and Sra, S. Can contrastive learning avoid shortcut solutions? *arXiv preprint arXiv:2106.11230*, 2021.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020a.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020b.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33: 9573–9585, 2020.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2164–2173. PMLR, 2019.

Stephenson, C., Padhy, S., Ganesh, A., Hui, Y., Tang, H., and Chung, S. On the geometry of generalization and memorization in deep neural networks. *arXiv preprint arXiv:2105.14602*, 2021.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021.

Tsai, Y.-H. H., Wu, Y., Salakhutdinov, R., and Morency, L.-P. Demystifying self-supervised learning: An information-theoretical framework. *arXiv e-prints*, pp. arXiv–2006, 2020.

Tsai, Y.-H. H., Ma, M. Q., Zhao, H., Zhang, K., Morency, L.-P., and Salakhutdinov, R. Conditional contrastive learning: Removing undesirable information in self-supervised representations. *arXiv preprint arXiv:2106.02866*, 2021.

Tu, L., Lalwani, G., Gella, S., and He, H. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020.

Wang, Z. and Culotta, A. Identifying spurious correlations for robust text classification. *arXiv preprint arXiv:2010.02458*, 2020.

Xie, Q., Dai, Z., Du, Y., Hovy, E., and Neubig, G. Controllable invariance through adversarial feature learning. *arXiv preprint arXiv:1705.11122*, 2017.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhang, C., Bengio, S., Hardt, M., Mozer, M. C., and Singer, Y. Identity crisis: Memorization and generalization under extreme overparameterization. *arXiv preprint arXiv:1902.04698*, 2019.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

Zhao, H. and Gordon, G. J. Inherent tradeoffs in learning fair representations. *arXiv preprint arXiv:1906.08386*, 2019.

Zhou, H., Vani, A., Larochelle, H., and Courville, A. Fortuitous forgetting in connectionist networks. *arXiv preprint arXiv:2202.00155*, 2022.

A. Related Work

A.1. Group Robustness

Empirical risk minimization (ERM) Empirical risk minimization minimizes the average training loss across training points. Given a loss function $\ell(x, y; \theta) : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}_+$ (e.g. cross-entropy loss), ERM minimizes the following objective:

$$J_{\text{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta). \tag{1}$$

Group distributionally robust optimization (Group DRO) Group DRO uses training group annotations to minimize the worst-group error on the training set, assuming we have access to group annotations on the training data such that the n training points are $\{(x_1, y_1, g_1), \dots (x_n, y_n, g_n)\}$. For some loss function $\ell(x, y; \theta)$, the group DRO objective can then be written as:

$$J_{\text{group-DRO}}(\theta) = \max_{g \in \mathcal{G}} \frac{1}{n_g} \sum_{i | g_i = g} \ell(x_i, y_i; \theta) \tag{2}$$

where n_g is the number of training points with group $g_i = g$.

Just Train Twice (JTT) JTT is a simple two-stage approach that does not require group annotations at training time. First, it trains an identification model $\hat{f}_{\text{id}}$ via ERM and then identifies an error set $E = \{(x_i, y_i) \text{ s.t. } \hat{f}_{\text{id}}(x_i) \neq y_i\}$ of training examples that $\hat{f}_{\text{id}}$ misclassifies. Then, it trains a final model $\hat{f}_{\text{final}}$ by upweighting the points in the identified error set.

$$J_{\text{up-ERM}}(\theta, E) = \left(\lambda_{\text{up}} \sum_{(x,y) \in E} \ell(x, y; \theta) + \sum_{(x,y) \notin E} \ell(x, y; \theta) \right), \tag{3}$$

Correct-n-Contrast (CnC)

A.2. Self-supervised Representation Learning

Self-supervised representation learning methods learn visual representations from large-scale unlabeled images where data annotations are scarce and time-consuming. Contrastive learning is a discriminative approach to learn representations that aims to attract similar or positive samples and push apart different or negative samples, which has become increasingly successful in recent years (Chen et al., 2020a; He et al., 2019; Grill et al., 2020; Caron et al., 2020; **?**). The standard approach for generating positive pairs without additional annotations is to create multiple views of each data point using random augmentations. The contrastive learning loss or InfoNCE (Oord et al., 2018) then maximizes a lower bound on the mutual information between the two views.

For instance, in MoCo (He et al., 2019), a dictionary lookup task is defined in which for each encoded query q, there is a set of encoded keys $\{k_0, k_1, \dots\}$, among which a single positive key k_+ matches the query q, which is generated from different views. The contrastive loss function is utilized to pull q close to q_+ while pushing it away from other negative keys:

$$L_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\exp(q \cdot k_+ / \tau) + \sum_{k_-} \exp(q \cdot k_- / \tau)}$$

SimCLR (Chen et al., 2020a) employs the same loss function but samples negative pairs from the current batch while MoCo uses a momentum memory bank. More specifically, SimCLR generates two randomly augmented views of each image $\tilde{\boldsymbol{x}}_i = t(\boldsymbol{x}), \tilde{\boldsymbol{x}}_j = t'(\boldsymbol{x}), \quad t, t' \sim \mathcal{T}$ given a batch of images, and uses all other augmentated samples from the batch as negative examples. Then it uses a neural encoder $f(\cdot)$ to extract representations from these augmented examples, and a small projection head $g(\cdot)$ which maps these representations to the contrastive loss space. More specifically, the loss function on a

minibatch of N samples would be as below.

$$\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}[\ell(2k-1,2k)+\ell(2k,2k-1)]$$
$$\ell(i,j) = -\log\frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} k \neq i \exp(s_{i,k}/\tau)},$$
$$s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j/(\|\boldsymbol{z}_i\|\|\boldsymbol{z}_j\|)$$

More recent methods discard the need for negative samples in contrastive learning. BYOL (Grill et al., 2020) uses a siamese architecture with momentum encoders to prevent different representations from collapsing into one vector. SwAV (Caron et al., 2020) exploits online clustering for each batch to enforce consistency between cluster assignments from different views, and SimSiam (Chen & He, 2020) uses a simple stop-gradient operation in a siamese architecture to avoid collapsing.

Theoretical analysis (Arora et al., 2019; Tsai et al., 2020; Tosh et al., 2021) suggests the contrastively learned representations can lead to a good downstream performance. Moreover, Tian et al. (2020) shows what matters more for the performance are the choice of the contrastive learning objective, and the creation of the positive and negative data pairs in the contrastive objective. Recent work (Khosla et al., 2020) extends the usage of contrastive learning from the self-supervised setting to the supervised setting. The supervised setting defines the positive pairs as the data from the same class in the contrastive objective, while the self-supervised setting defines the positive pairs as the data with different augmentations.

A.3. Fair Representation Learning

A family of fairness definitions for SSL have been previously explored in representation learning (Edwards & Storkey, 2015; Beutel et al., 2017; Louizos et al., 2015; Madras et al., 2018). In this setting, they train representations on data includeing sensitive attributes, such that classifiers trained on these representations are more likely to be agnostic to the these attributes.

The works in fair representation learning usually address removing sensitive attributes from the representations by: obfuscating any information about sensitive attributes in order to approximately satisfy demographic parity (Zemel et al., 2013), using adversarial methods (Edwards & Storkey, 2015; Xie et al., 2017; Beutel et al., 2017; Zhang et al., 2018; Madras et al., 2018; Adel et al., 2019; Zhao & Gordon, 2019), or feature disentanglement based using variational approaches. (Kingma & Welling, 2013; Gretton et al., 2006; Louizos et al., 2015; Amini et al., 2019; Alemi et al., 2018; Burgess et al., 2018; Chen et al., 2018; Kim & Mnih, 2018; Esmaeili et al., 2019).

A.3.1. Fair Self-supervised Representation Learning

The pioneer work of Song et al. (2019) proposes an information theoretically motivated objective for learning maximally expressive representations subject to fairness constraints. They try to maximize the expressiveness of representations while satisfying certain fairness constraints, or in other words to represent expressiveness as well as three dominant notions of fairness (demographic parity, equalized odds, equalized opportunity) in terms of mutual information, and obtain tractable upper/lower bounds of these mutual information objectives.

A group of recent works in contrastive representation learning focus on proposing models that explicitly tackle the problem of imbalanced and long-tailed distribution real-world data. Jiang et al. (2021b) have proposed a framework in which they create a dynamic self-competitor model to contrast with the target model, which is a pruned version of the target model. During training, contrasting the two models will lead to adaptive online mining of the most easily forgotten samples for the current target model, and implicitly emphasize them more in the contrastive loss.

Another group of work focuses on removing undesirable information in self-supervised representations. They focus on datasets that certain information that may be undesirable for downstream tasks. For instance, gender information which may lead to biased decisions on many gender-irrelevant tasks. (Tsai et al., 2021) have developed conditional contrastive learning to remove the effect of the undesirable variable which conditions on the undesirable variable (by fixing the variations of it) during the contrastive learning process. In particular, they introduce Conditional InfoNCE (C-InfoNCE), and its computationally efficient variant, Weak-Conditional InfoNCE (WeaC-InfoNCE), for conditional contrastive learning.

B. Datasets

We make use of the following four image datasets:

- `waterbirds` (Sagawa et al., 2020a): Background (land, water) spuriously correlated with bird (labdbird, waterbird)
- `cmnist` (Colored MNIST) (Arjovsky et al., 2019): Color of digit on the images spuriously correlated with the number
- `spurcifar10` (Spurious CIFAR10) (Nagarajan et al., 2020): Color of lines on the images spuriously correlated with the class
- `metashift` (Liang & Zou, 2022): Cats vs Dogs task: Background (indoor, outdoor) spuriously correlated with pet (cat, dog)

C. Comparing SSL to CLIP representations

We train linear classifiers with different re-sampled sets of training examples. As shown in table **??**, even CLIP representations do not help mitigate the geometrical and statistical skews when learning the linear classifier on representations.

		k-NN		Linear probe		Spurious Attribute (Linear)	
dataset	Linear Train set	Average	Worst-group	Average	Worst-group	Average	Worst-group
CelebA	Normal	83.51%	20.39%	89.67%	16.98%	77.09%	61.26%
	Downsampled	78.57%	75.56%	88.83%	81.11%	97.04%	91.11%
	Upsampled	90.89%	28.33%	91.76%	86.11%	98.57%	90.56%
Waterbirds	Normal	56.77%	30.53%	61.30%	45.08%	68.06%	45.09%
	Downsampled	69.23%	63.81%	79.01%	74.01%	89.59%	86.30%
	Upsampled	74.06%	43.93%	75.35%	55.61%	83.26%	68.85%
Metashift	Normal	74.44%	50.85%	76.06%	7.91%	63.19%	22.60%
	Downsampled	84.69%	68.97%	80.69%	30.51%	73.68%	22.03%
	Upsampled	88.41%	73.79%	82.55%	32.20%	75.54%	35.59%
Colored MNIST	Normal	73.59%	54.65%	72.64%	55.57%	74.14%	70.40%
	Downsampled	91.84%	89.39%	89.04%	85.98%	98.08%	96.21%
	Upsampled	97.18%	62.88%	93.76%	88.26%	97.95%	97.16%
Spurious CIFAR10	Normal	25.11%	14.58%	35.79%	22.87%	93.62%	0.00%
	Downsampled	33.20%	10.42%	52.56%	33.33%	51.29%	30.00%
	Upsampled	39.28%	20.83%	58.17%	41.67%	58.14%	35.42%

Table 3: Average and worst-group test accuracies of CLIP representations trained on Normal, , Downsampled, Upsampled-train sets. Balancing the training set used for linear evaluation helps us identify the learned representations by avoiding the statistical and geometrical skews

D. Additional Representation Augmentation Results

Even without having access to group information in the validation set, untuned hyperparameters (layer threshold and pruning percentage), we improve worst-group accuracy in most cases.

		kNN (Target)		LR (Target)		LR (Group)	Other Metrics		
Dataset	**Method**	**Worst Group**	**Average**	**Worst Group**	**Average**	**Average**	**I(Z; Y)**	**I(Z; G)**	$\mathbf{L_{align}}$
cmnist	SSL_{Base}	44.4%	85.6%	39.0%	89.8%	84.5%	8.59E-01	8.82E-05	1.129
	$\text{SSL}_{\text{Reinit, 17}}$	39.0%	81.2%	35.6%	87.8%	81.3%	8.48E-01	8.76E-01	0.620
	$\text{SSL}_{\text{Reinit, 19}}$	35.6%	80.3%	37.3%	89.0%	82.0%	8.57E-01	9.05E-01	0.875
	$\text{SSL}_{\text{Thres, 0.7}}$	37.3%	84.3%	39.0%	88.8%	83.2%	8.58E-01	9.06E-01	1.235
	$\text{SSL}_{\text{Thres, 0.95}}$	33.3%	84.6%	31.5%	90.9%	84.3%	8.53E-01	7.19E-01	1.141
metashift	SSL_{Base}	18.6%	65.5%	31.0%	57.5%	62.9%	8.41E-02	8.43E-02	0.925
	$\text{SSL}_{\text{Reinit, 17}}$	16.9%	62.8%	26.2%	57.2%	64.5%	7.37E-06	9.31E-02	0.906
	$\text{SSL}_{\text{Reinit, 19}}$	15.3%	67.0%	23.4%	60.8%	65.2%	1.19E-02	1.42E-01	0.942
	$\text{SSL}_{\text{Thres, 0.7}}$	15.3%	64.7%	22.8%	61.4%	65.2%	1.05E-02	1.69E-01	0.971
	$\text{SSL}_{\text{Thres, 0.95}}$	11.9%	62.8%	27.6%	58.5%	64.9%	5.21E-09	3.05E-02	0.924
spurcifar10	SSL_{Base}	27.1%	54.7%	11.1%	44.0%	61.7%	1.51E+00	4.68E-03	0.750
	$\text{SSL}_{\text{Reinit, 17}}$	31.7%	51.5%	25.8%	48.8%	50.2%	1.28E+00	4.63E-03	0.333
	$\text{SSL}_{\text{Reinit, 19}}$	31.3%	59.3%	20.8%	54.8%	49.5%	1.83E+00	6.86E-10	0.644
	$\text{SSL}_{\text{Thres, 0.7}}$	31.3%	57.7%	29.2%	51.5%	47.9%	1.88E+00	8.12E-11	0.946
	$\text{SSL}_{\text{Thres, 0.95}}$	38.5%	59.0%	30.6%	54.0%	48.6%	1.89E+00	1.47E-10	0.894
waterbirds	SSL_{Base}	43.4%	50.2%	46.2%	49.5%	84.4%	8.23E-03	5.41E-01	0.706
	$\text{SSL}_{\text{Reinit, 17}}$	45.5%	50.0%	46.8%	51.6%	84.5%	5.78E-02	4.31E-01	0.282
	$\text{SSL}_{\text{Reinit, 19}}$	46.7%	55.2%	22.7%	55.6%	88.3%	5.12E-02	5.82E-01	0.392
	$\text{SSL}_{\text{Thres, 0.7}}$	51.6%	62.1%	43.8%	55.0%	81.8%	7.86E-02	5.72E-01	0.873
	$\text{SSL}_{\text{Thres, 0.95}}$	42.2%	58.0%	38.5%	53.6%	87.1%	9.69E-02	5.39E-01	0.684

Table 4: The same set of augmentations for each dataset; without using group annotations of validation data, worst-group accuracy can be improved.