Large Learning Rates Simultaneously Achieve Robustness to Spurious Correlations and Compressibility

Melih Barsbey Lucas Prieto Stefanos Zafeiriou Tolga Birdal
Imperial College London

Abstract

Robustness and resource-efficiency are two highly desirable properties for modern machine learning models. However, achieving them jointly remains a challenge. In this paper, we identify high learning rates as a facilitator for simultaneously achieving robustness to spurious correlations and network compressibility. We demonstrate that large learning rates also produce desirable representation properties such as invariant feature utilization, class separation, and activation sparsity. Our findings indicate that large learning rates compare favorably to other hyperparameters and regularization methods, in consistently satisfying these properties in tandem. In addition to demonstrating the positive effect of large learning rates across diverse spurious correlation datasets, models, and optimizers, we also present strong evidence that the previously documented success of large learning rates in standard classification tasks is related to addressing hidden/rare spurious correlations in the training dataset. Our investigation of the mechanisms underlying this phenomenon reveals the importance of confident mispredictions of bias-conflicting samples under large learning rates.

1. Introduction

The requirement to function well in novel circumstances and being resource-efficient are two central challenges for modern machine learning (ML) systems, which are expected to perform critical functions with less resources and on smaller hardware, in the face of limited access to computational resources, environmental concerns about energy consumption, and tightening bottlenecks around computational hardware [9, 10, 21]. As ML-based technologies increasingly permeate every aspect of daily and industrial life [25], it becomes urgent to determine the inductive biases that help the models address both challenges together.

Although no learner can be expected to perform well under arbitrary changes in the environment [80], they are expected to do so under "reasonable" distribution shifts – a capability commonly exhibited by numerous animal species.

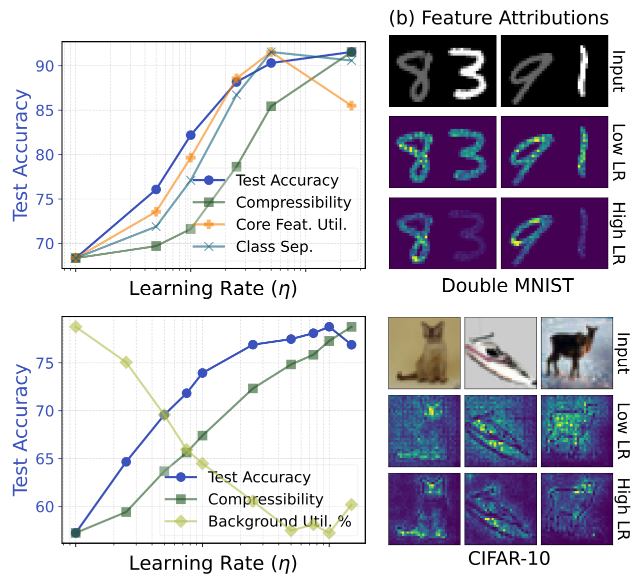


Figure 1. (**Top left**) When trained on the Double MNIST spurious correlation (SC) dataset with higher learning rates (LR), models tend to be more robust against SCs (higher accuracy), more compressible, and have more favorable core/invariant feature utilization and class separation. (**Top right**) Given images containing core and spurious features (bright vs. dim digit), high LR models are more likely to be attuned to the core feature vs. spurious feature, unlike low LR models. (**Bottom**) Our findings extend to standard classification tasks since vulnerability to SCs, such as background, strongly associate with accuracy drops. All ranges normalized in y-axis to highlight relationship to LR.

Such robustness to distributions that differ from the training data in principled ways have been studied under the umbrella term out-of-distribution (OOD) generalization [79]. Although scaling models and datasets, as well as targeting specific aspects of this issue have shown some success [6, 45], OOD generalization remains an open problem both theoretically and in practice [12].

A critical obstacle for OOD generalization is the presence of spurious correlations[1] (SCs) in training data, which can intuitively be described as the relationships between the

[1] Unless noted otherwise, throughout the paper any references to robustness or OOD generalization will be in the context of spurious correlations.

input features and output label in the training set that do not transfer to the test set. A canonical example is highlighted by [4], where models are likely to misclassify a cow on sand as a camel, due to learning the misleading statistical association between camels and desert backgrounds in the training data. [65] show empirically and theoretically how overparameterization can cause pathological learning dynamics, amplifying the emphasis on SCs and degrading OOD performance. Similarly, [17, 37] show that AI systems and large language models (LLMs) can be susceptible to large performance drops if training data contains easily exploitable patterns not carrying over to new environments.

The challenge of OOD generalization becomes even more pronounced when coupled with the need for model compressibility, given the increased demands of resource-efficiency on modern ML systems. It remains unclear whether these two objectives –robustness and efficiency– are inherently in conflict or can actually complement each other [14, 18]. We posit that understanding this interplay is critical for the design of next generation ML models.

To date, explaining the counterintuitive observation that overparameterization often improves rather than harms generalization, especially under SCs, remains to be illuminated, as does understanding when and why models become compressible [2, 3, 5, 15, 55, 61]. One key factor in this puzzle is the learning rate (LR), or step size, used during gradient-based training. Various studies have highlighted the impact of large LRs on generalization [40, 43, 53], model compressibility [3], and representation sparsity [1]. Recent research confirms that LR remains a crucial hyperparameter in the training of multimodal foundational models [75].

In light of these, in this paper we hypothesize that *in deep neural networks, learning rate plays a pivotal role in achieving robustness to SCs without sacrificing efficiency*. We then confirm this hypothesis via extensive analyses, making the following key contributions:

1. We establish that **large LRs simultaneously and consistently promote both compressibility and robustness** specifically to SCs across a wide range of architectures, datasets, and training schemes.
2. We identify that these effects are accompanied by **improved core feature utilization, class separation**, and **compressibility** in the learned representations.
3. We show that large LRs produce **a unique combination of the aforementioned desirable properties**, in comparison to other major hyperparameters and regularizers.
4. We provide strong evidence that the robustness against SCs that large LRs confer, **contribute** to their previously documented success in **standard generalization tasks**.
5. Our investigation into the mechanisms reveal the importance of **confident mispredictions of bias-conflicting samples** under high LRs.

In Fig. 1 we share an overview of our results.

2. Related Work

Robustness to Spurious Correlations. Overreliance on simple, easily exploitable features that have limited bearing on a test set have been pointed out by [22] and [67]. [65] examine the role of overparametrization in producing models that rely on spurious features, and [54] point out how two features of the data distribution (that they name geometrical and statistical skews) might lead to a max-margin classifier ending up utilizing spurious features. [57] point out the importance of features learned in early training, where easy-to-learn, spurious features might not be replaced by better-generalizing features. Various methods have been previously proposed to alleviate this problem. Methods that assume access to spurious feature labels/annotations exploit this information in different ways to improve worst group or unbiased test set performance [28, 64]. In the absence of group annotations, alternative methods rely on assumptions about the nature of the spurious features and the inductive biases of the learning algorithms [45, 59, 74].

Inductive Bias of Large Learning Rates, Model Compressibility. [43] provide one of the earliest set of findings regarding the inductive bias of LRs in standard machine learning tasks, and examine how large vs. small LRs lead to qualitatively different features to be learned by the neural network. [29] point out how LRs in early training prevents the iterates from being locked into narrow valleys in the loss landscape, where the curvature in certain directions are high, to the detriment of the conditioning of the gradient covariance matrix. [40, 53] point out the importance of large LRs in early training [30]. [63] demonstrates the crucial role of spurious / opposing signals in early training, and how progressive sharpening [8, 78] of the loss landscape in the directions that pertain to the representation of these features lead to the eventual down-weighting of such non-robust features. [14] find that lottery-ticket style pruning methods present the most advantageous trade-off between performance, robustness, and compressibility. To date, no study has systematically explored how LR influences robustness to SCs or how it interacts with compressibility. See our suppl. material for an extended literature review.

3. Setup

We consider a classification setting, where the task of a model (e.g.,, a neural network) f is to predict the discrete label $y \in \mathcal{Y}$, given a d-dimensional input $\mathbf{x} \in \mathbb{R}^d$. In order to assess the quality of a neural network represented by its weights $\mathbf{w}$, we consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$, such that $\ell(y, f_{\mathbf{w}}(\mathbf{x}))$ measures the error incurred by predicting the label of $\mathbf{x}$ as $\arg\max_j f_{\mathbf{w}}(\mathbf{x})[j]$, when the true label is y. We define a *data distribution* μ_Z over $\mathcal{Z}$, and a *training dataset* with n elements, *i.e.*,, $S = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, where each $\mathbf{z}_i := (\mathbf{x}_i, y_i) \overset{\text{i.i.d.}}{\sim} \mu_Z$. We then denote the *population* and *empirical risks* as $\mathcal{R}(\mathbf{w}) := \mathbb{E}_{\mathbf{x},y}\left[\ell(y, f_{\mathbf{w}}(\mathbf{x}))\right]$

and $\widehat{\mathcal{R}}(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f_{\mathbf{w}}(\mathbf{x}_i))$. Unless otherwise noted, we utilize the (minibatch) stochastic gradient descent (SGD) algorithm for empirical risk minimization: $\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\nabla_{\mathbf{w}^t}\frac{1}{b}\sum_{i\in\Omega^t}\ell(y_i, f_{\mathbf{w}}(\mathbf{x}_i))$, where Ω^t is a randomly sampled fixed-size subset of the training set, $b := |\Omega^t|$ is batch size, and η is learning rate (LR; aka step size).

3.1. Spurious Correlations

OOD generalization describes the case whenever we have access to $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, $\mathbf{z}_i := (\mathbf{x}_i, y_i) \overset{\text{i.i.d.}}{\sim} \mu_Z^{\text{train}}$, yet we are interested in evaluating risk under a different distribution $\mathcal{R}(\mathbf{w}) := \mathbb{E}_{\mathbf{x},y\sim\mu_Z^{\text{test}}}[\ell(y, f_{\mathbf{w}}(\mathbf{x}))]$, where $\mu_Z^{\text{train}} \neq \mu_Z^{\text{test}}$. This difference between training and test distributions is called *distribution shift*. Some examples of distribution shift include *subpopulation shift*, where $\mu_Z^{\text{train}}(y) \neq \mu_Z^{\text{test}}(y)$, and *mechanism shift*, where $\mu_Z^{\text{train}}(\mathbf{x}|y) \neq \mu_Z^{\text{test}}(\mathbf{x}|y)$. A subtype of mechanism shift is of special importance in the literature and of particular importance for this paper: A model trained on S is said to potentially be vulnerable to *spurious correlations*[2] in the training dataset when it is possible to define a decomposition of $\mathbf{x} = (\mathbf{x}^c, \mathbf{x}^s)$, such that $\mu_Z^{\text{train}}(\mathbf{x}^c|y) = \mu_Z^{\text{test}}(\mathbf{x}^c|y)$, and $\mu_Z^{\text{train}}(\mathbf{x}^s|y) \neq \mu_Z^{\text{test}}(\mathbf{x}^s|y)$. In such cases, $\mathbf{x}^c$ are frequently called *core features* (aka invariant features), and $\mathbf{x}^s$ are frequently called *spurious features*.

In previous research that addressed this problem, the assumed data distributions can vary considerably [56, 79]. In this study we focus on one of the most canonical cases, previously called "perception tasks" [59] or "easy-to-learn" tasks [54], where the core features are perfectly informative with respect to the label in the training set, and spurious features imperfectly so. In such tasks, the former are construed to be more difficult to learn. We assume a separate generative model for spurious features. We define a bias label $b \in \mathcal{Y}$, as opposed to the class label $y \in \mathcal{Y}$ for core features, admitting the decomposition $\mu_Z(\mathbf{x}^c, \mathbf{x}^s, y, b) = \mu_Z(\mathbf{x}^c|y)\mu_Z(\mathbf{x}^s|b)\mu_Z(y, b)$, and while $\mu_Z^{\text{test}}(y, b) = \mu_Z^{\text{test}}(y)\mu_Z^{\text{test}}(b)$, we have $\mu_Z^{\text{train}}(y, b) \neq \mu_Z^{\text{train}}(y)\mu_Z^{\text{train}}(b)$. More specifically, we further assume $\mu_Z^{\text{train}}(y = a, b = a) \gg \mu_Z^{\text{train}}(y = a)\mu_Z^{\text{train}}(b = a)$, where the mutual information between y and b in the training set presents a challenge for the learner. The value $\rho^{\text{train}} := 1 - \mu_Z^{\text{train}}(y = b)$ determines the rate of *bias-conflicting* examples in the training dataset, which a learner can exploit to avoid utilizing spurious features. Although there does not exist a canonical definition of easy vs. difficult-to-learn features in the literature [60], they can be construed as the difficulty of estimating $\mu_Z^{\text{train}}(y|\mathbf{x}^c)$ vs. $\mu_Z^{\text{train}}(b|\mathbf{x}^s)$.

[2]As is common in the literature, here we use the term spurious *correlations* to also include non-linear relationships induced by a confounder.

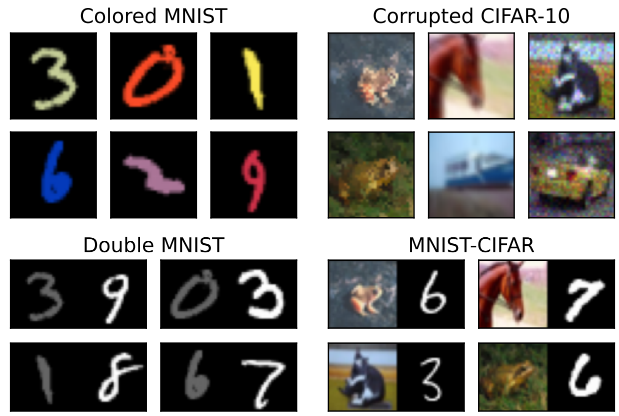


Figure 2. Example images from semi-synthetic SC datasets.

3.2. Compressibility

A frequently used metric for characterizing compressibility, especially for activations, is *sparsity*. Given a d-dimensional vector $\mathbf{z} \in \mathbb{R}^d$ it can be defined as $\text{Sparsity}(\mathbf{z}) = 1 - \|\mathbf{z}\|_0/d$. Although this is an intuitive and useful metric, more nuanced notions of compressibility are desirable, e.g. when considering activation functions like sigma that does not output 0 values. This is because a (deterministic or probabilistic) vector might be "summarizable" with a small subset of its elements regardless of the number of entries in the vector that *exactly* equal 0. Therefore, we use (q, κ)-Compressibility inspired by [23]: $(q, \kappa)\text{−Compressibility}(\mathbf{z}) = 1 - \frac{\inf_{\|\mathbf{y}\|_0 \leq \lceil \kappa d \rceil}\|\mathbf{z}-\mathbf{y}\|_q}{\|\mathbf{z}\|_q}$. Intuitively, if a vector's $(2, 0.1)$-Compressibility is high, this means that this vector can be approximated with little error (in an ℓ_2 sense) by using 0.1 of its d elements.

While sparsity or (q, κ)-Compressibility have merit as compressibility metrics, given our interest in compressibility *in relation to* OOD generalization performance, quantifying network/parameter compressibility while taking the downstream effects of compression on performance into account is crucial. For this purpose we define κ-Prunability; given a predictor f and its (unbiased) test accuracy $\text{Acc}_{\mu^{\text{test}}}(f)$ it is defined as $\kappa\text{P}(f) = \text{Acc}_{\mu^{\text{test}}}(f^{(\kappa)}) \,/\, \text{Acc}_{\mu^{\text{test}}}(f)$, where $f^{(\kappa)}$ corresponds to f with $\kappa \in [0, 1]$ of its parameters pruned (set to 0). This measures the retention of (unbiased) test accuracy after pruning a fraction κ of parameters. Here, we utilize structured (neuron/filter) pruning given its computationally desirable properties [41]. Instead of committing to a particular κ, we will use $\sum_{\kappa\in\mathbf{k}}\kappa\text{P}(f)$, $\mathbf{k} = (0.1, 0.2, \ldots, 0.9)$. Given our chosen metric, we will use the terms network prunability and network compressibility interchangeably throughout the rest of the text. See suppl. material for qualitatively identical results with alternative metrics.

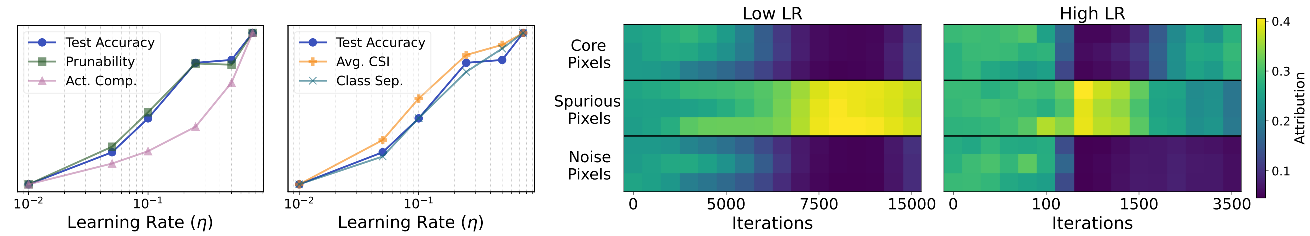


Figure 3. (Left) Effects of learning rate on OOD performance (unbiased test acc.), network prunability, and representation properties with the parity dataset. See suppl. material for min. and max. values. (Right) Prediction attributions (i.e. feature importance) for core, spurious, and noise pixels throughout training for low and high LR models.

3.3. Metrics for Representation Analysis

We now define metrics we use when analyzing learned representations beyond compressibility. Unless otherwise noted, all analyses on representations target post-activation values in the penultimate neural network layer, often referred to as *learned representations* in the literature.

Class Separation. We use different metrics to quantify representations' class-sensitivity. The first metric we utilize is class separation R^2, where we adopt the definition of [34], who investigate this notion in relation to model generalization performance as well as representations' transferability under different training losses: $R^2 = 1 - \bar{d}_{\text{within}}/\bar{d}_{\text{total}}$, where $\bar{d}_{\text{within}} = \sum_{k=1}^{K} \sum_{m=1}^{N_k} \sum_{n=1}^{N_k} \frac{1-\text{sim}(x_{k,m}, x_{k,n})}{K N_k^2}$ and $\bar{d}_{\text{total}} = \sum_{k=1}^{K} \sum_{j=1}^{K} \sum_{m=1}^{N_j} \sum_{n=1}^{N_k} \frac{1-\text{sim}(x_{j,m}, x_{k,n})}{K^2 N_j N_k}$, with K denoting number of classes, N_j number of samples with label j, and sim a similarity metric such as cosine similarity.

Core vs. Spurious Feature Utilization. Though important for a classification task, class separation does not quantify sensitivity of specific neurons to classes. For this, we use *class-selectivity index* (CSI) [38, 62]: CSI $= \rho_{\max} \frac{\pi_{\max} - \pi_{-\max}}{\pi_{\max} + \pi_{-\max} + \epsilon}$. As in prior works, we take $\pi_{\max}$ to be largest class-conditional activation mean for a given neuron, where the means are computed over a sample of inputs, and $\pi_{-\max}$ is the mean of the remaining classes. In contrast to previous use of this metric, given the prevalence of activation sparsity in our experiments, we scale this fraction with $\rho_{\max}$, which corresponds to the ratio of instances belonging to the said class for which the neuron is activated. This factors in the "coverage" of a particular neuron for the instances belonging to a class, and prevents computing high class-selectivity for a neuron that fires rarely for the members of the said class. When computed over an unbiased test set, this metric serves to characterize a neuron's *sensitivity to core features*. This is especially useful when the core features and spurious features overlap and attribution methods cannot be reliably used to investigate a specific neuron's responsiveness to core features. We average over neurons' class sensitivities (in learned representations) to characterize a network's core feature utilization. Note that if accessible, spurious feature labels can be used to compute the counterpart bias-selectivity index (BSI).

Input Image Attributions. To quantize the influence of particular input pixels or features over models' predictions, we utilize the commonly used interpretability method Integrated Gradients (IG) [71] to compute pixel-wise attributions over 10 random seeds. See suppl. material for details of IG, identical results under other interpretability methods, and how it provides convergent results with the aforementioned CSI regarding core/spurious feature utilization.

4. Datasets, Models, and Training Procedure

4.1. Datasets

We investigate the effect of LR on the model behavior on four classes of datasets: 1- Synthetic SC, 2- Semi-synthetic SC, 3- Naturalistic SC, and 4- Naturalistic classification. We describe each class of datasets below. As described in our setup, all SC datasets will involve a simple (spurious) feature that is predictive of the true label in the training set but not in the test set, and a more complex (core) feature that is predictive of the label in both training and test.

Synthetic SC data. We utilize two synthetic datasets from the literature to investigate the effects of LR on robustness to SCs and compressibility. These are the *parity dataset* proposed in [60] and the *moon-star dataset* proposed in [22]. The advantage of utilizing synthetic datasets is the clear definition of *simple* vs. *complex* features they enable, as well as total control they afford over bias-conflicting sample ratio (ρ). In the parity dataset, the core and spurious features are binary vectors of size C and S, with their parity bits (i.e. whether the vectors include odd number of 1's) corresponding to the true label y vs. spurious label b. Setting $C > S$ leads the spurious feature to be simpler than the complex feature. The moon-star dataset is a binary classification task where the classifier is expected to distinguish moon shaped objects from star shaped objects, with the spurious feature being the quadrant of the image on which the object is located. For both datasets we set bias-conflicting sample ratios as $\rho^{\text{train}} = 0.1$ and $\rho^{\text{test}} = 0.5$.

Semi-synthetic SC data. These are arguably the most commonly used dataset types in research on SCs [32, 56, 59, 60, 65]. The reason for their popularity is the attractive combination they provide in the form of having relatively realistic inputs with a decent control over dataset proper-

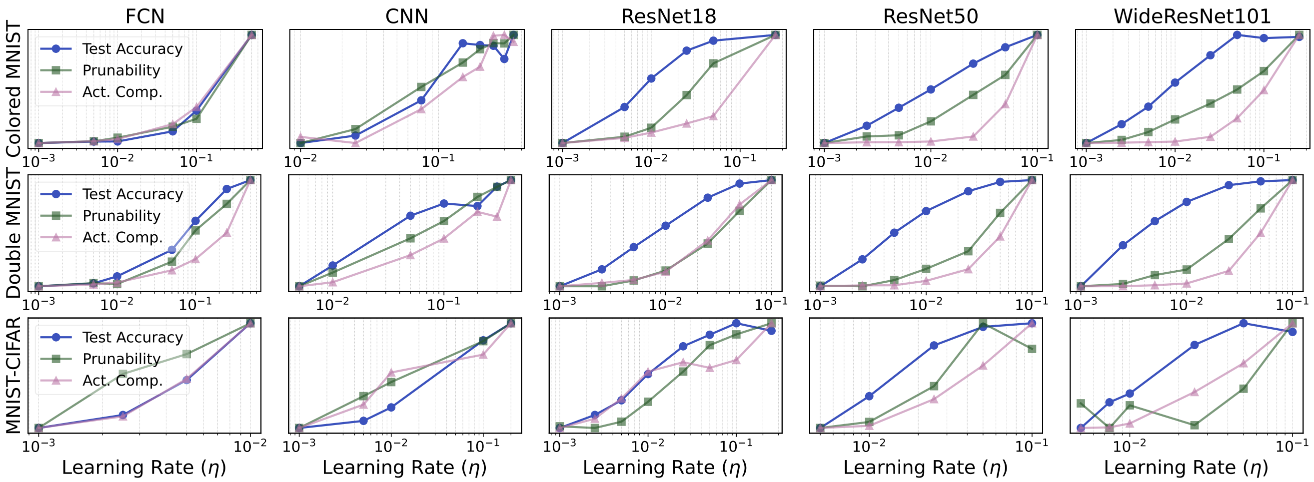


Figure 4. Effects of learning rate on OOD performance (unbiased test acc.), network prunability, and representation (activation) compressibility in semi-synthetic SC data. y-axes are normalized within each figure for each variable, see suppl. material for min. and max. values.

ties such as bias-conflicting sample ratios and the complexity of the core vs. spurious features. We present the bulk of our results on four such datasets: Colored MNIST, Corrupted CIFAR-10, MNIST-CIFAR, and Double MNIST (see Fig. 2 for examples from each). The former two were proposed by [56], and have been frequently used in the literature. MNIST-CIFAR dataset is the extension of the domino dataset proposed by [67] to all 10 classes of MNIST and CIFAR-10, while Double MNIST has been designed by the authors for this paper. Each dataset includes some combination and/or modification of the well-known image datasets MNIST [39] and CIFAR-10 [35]. The core and spurious features for these datasets can be specified as digit shape vs. digit color for Colored MNIST, object vs. corruption type for Corrupted CIFAR-10, left digit vs. (brighter) right digit for Double MNIST, and CIFAR-10 vs. MNIST targets for MNIST-CIFAR. We set $\rho^{\text{train}} = 0.025$ for Colored MNIST and Double MNIST datasets, and $\rho^{\text{train}} = 0.1$ for the remaining two, given the higher baseline difficulty thereof.

Naturalistic SC data. We also investigate the effect of LR on two naturalistic image classification datasets CelebA [46] and Waterbirds [76], both of which have been frequently used in research on SCs [64]. In the CelebA dataset, spurious and core features are the hair color and gender of the person in the images respectively. In the Waterbirds dataset, these correspond to the background of the pictured bird and their natural habitat (water vs. land). In keeping with literature [83], we examine worst-group test accuracy in our experiments with CelebA and Waterbirds [64], while for the rest we use unbiased test set accuracy [56].

Standard classification data. As we extend our inquiry from SC datasets to naturalistic classification tasks, we investigate the effects of LR on model behavior using CIFAR-10, CIFAR-100, and ImageNet-1k datasets [11, 35]. The former two include 32×32 images of 10 and 100 classes

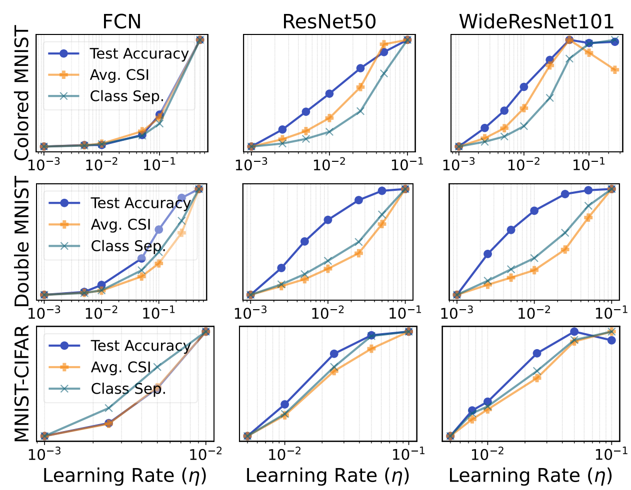


Figure 5. Effects of learning rate on OOD performance (unbiased test acc.) and representation properties in semi-synthetic SC datasets. y-axes are normalized within each figure for each variable, see suppl. material for min. and max. values.

of objects or animals, respectively. We use the traditional train-test splits in our experiments, consisting of 50000 and 10000 samples. ImageNet-1k consists of 1,331,167 color images scaled to 256×256 belonging to 1000 classes. We use the traditional train-validation split with 1,281,167 and 50000 samples each.

4.2. Architectures, Training, & Implementation

To investigate the diversity of contexts in which LR has the aforementioned effects, we utilize a variety of architectures. These include fully connected networks (FCN) with ReLU activation, convolutional neural networks (CNN; a slightly simplified variant of VGG11 [69]), ResNet18 and ResNet50 [26], and Wide ResNet-101-2 [84]. We further include a vision transformer model, namely the Swin Transformer by

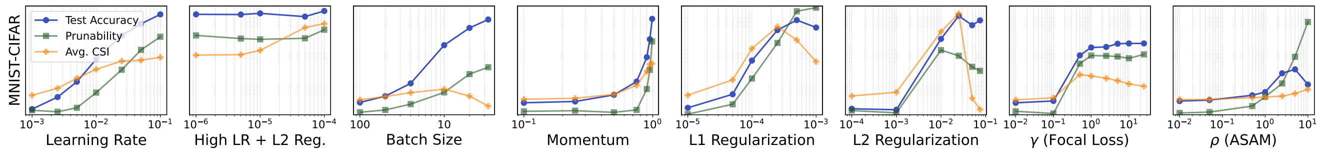


Figure 6. Comparing hyperparameters, regularization methods, and losses in terms of OOD robustness, compressibility, and core feature utilization. y-axes are comparable within variables across rows, see suppl. material for min. and max. values.

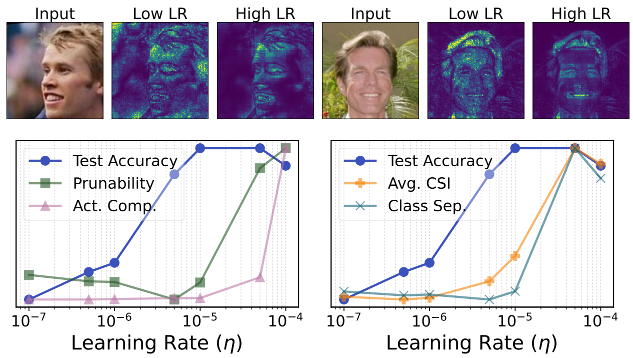


Figure 7. Effects of LR with a Swin Transformer and CelebA dataset on (top) model attributions and (bottom) network statistics.

[47]. Unless otherwise stated, all models are trained with a constant LR until 100% accuracy with no explicit L1/L2 regularization. Batch sizes are set to 16 for CelebA, Waterbirds, and ImageNet experiments, and 100 for the remaining experiments. An exception to the training scheme above is the training of the Swin Transformer on CelebA and Waterbirds datasets, where we utilize an AdamW optimizer combined with an ImageNet-1K pretraining initialization, to investigate the effects of LRs under more realistic training schemes. Our suppl. material provides additional details for our experiments and implementation, while showing that using different convergence criteria, optimizers (Adam [31], PSGD [42]), or LR schedulers (Cosine annealing [48], WSD [27]) leads to qualitatively identical, and sometimes stronger results. Unless otherwise noted, the results presented are the average of experiments run with at least 3 different random seeds. Since high LRs can lead to divergence and/or drastic performance loss at the extremes, we utilize LRs up to the point of divergence or drastic performance loss ($> 25\%$ in test accuracy) among any one of the seeds. Lastly, since our results involve the examination of a number of variables simultaneously, in figures y-axes are normalized for each variable to highlight the effects of LR, and value ranges for all variables are separately provided in the supp. material for readability. See our public repository[3] for additional implementation details.

5. Results

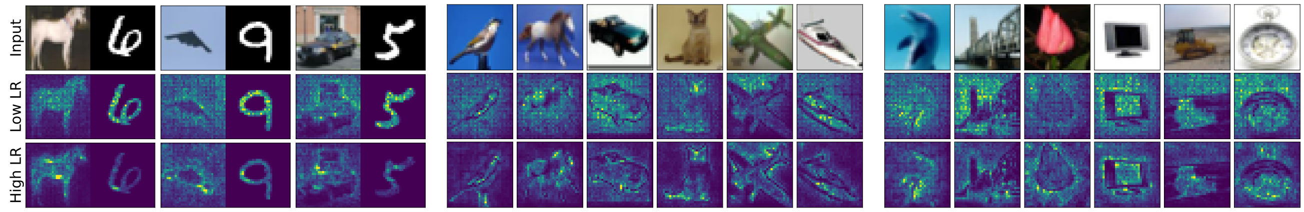
How does LRs affect robustness to SC and compressibility? We first investigate the effects of LR on compressibility and robustness to SCs, using the parity dataset in Fig. 3, with a fully connected network (FCN) with 3 hidden layers of width 200. The results show a clear effect of LR on robustness to SC and compressibility: Robustness and compressibility increases as a function of LR. The same figure also shows how the attributions to core, spurious, and noise pixels/features evolve for low and high LR models. Moving on to semi-synthetic SC datasets. Fig. 4 shows that across datasets and architectures, larger LRs lead to increased robustness to SCs, strongly supporting our central hypothesis. This increase in robustness is also accompanied by increased network and activation compressibility. Recall that the notion of κ-Prunability we employ here is computed under an unbiased test set, emphasizing the high LRs effect on achieving these aims jointly. See suppl. material for additional details and similar results with moon-star and Corrupted CIFAR-10 datasets.

How does LR affect learned representations? Our results regarding representation properties induced by high LRs are presented in Fig. 5. The results demonstrate that the increase in unbiased test performance and compressibility is accompanied by an increase in core feature utilization and class separation in the learned representations. Given recent results demonstrating learned representations' insensitivity to explicit SC interventions [32], it is important to highlight that the positive effects of high LRs are reflected in the learned representations themselves.

Do the observed effects of LR generalize to larger models? To make sure our results are not limited to small models and (semi-)synthetic data, we investigate the effect of LRs using a larger, transformer architecture and naturalistic data. As described in Sec. 4.2, we train a Swin Transformer on the CelebA dataset, and present its effects on worst-group performance, compressibility, and representation properties in Fig. 7. The results show that high LRs co-achieve robustness and compressibility in this scenario as well, further supporting our motivating hypothesis. See suppl. material for similar results on Waterbirds dataset.

How does LR behave when compared to and combined with other hyperparameters? We then move on to investigate how large LRs compare to other fundamental hyperparameters in a machine learning pipeline. The comparison set includes batch size, momentum, L1 and L2 regularization of the parameters, an alternative loss function designed for better exploitation of rare difficult examples (focal loss;

[3] https://github.com/mbarsbey/spurious_comp



(a) MNIST-CIFAR (b) CIFAR-10 (c) CIFAR-100

Figure 8. Attributions of a trained ResNet18 model on different datasets. For all datasets, top row includes original images; middle and bottom rows includes attributions of a model trained under low vs. high learning rate respectively.

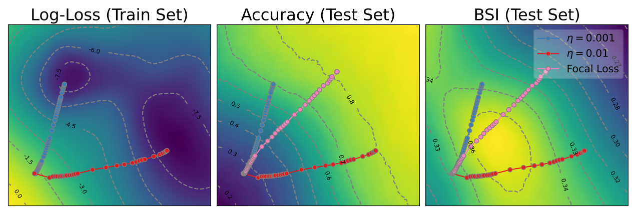


Figure 9. Training loss, unbiased test acc., and spurious feature utilization landscapes for models with different LRs and FL.

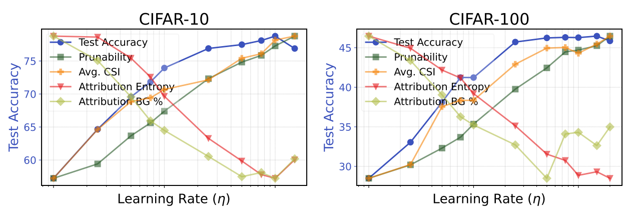


Figure 10. Higher LRs yield favorable generalization, compressibility, and core feature utilization for CIFAR datasets.

FL [44]), and an optimization procedure for finding well-generalizing (wide) minima in the training loss landscape (adaptive sharpness aware minimization; ASAM [19, 36]). We repeat our experiments by varying the each hyperparameter in question (In the case of FL and ASAM these correspond to their central hyperparameters respectively, γ and ρ). Fig. 6 demonstrates our results using a ResNet18 model: High LRs consistently facilitate a combination of OOD generalization, network compressibility, and core feature utilization, either on par with or surpassing other interventions. We also observe that other interventions (e.g. L1/L2 regularization) can be combined with high LRs to achieve even better trade-offs. Importantly, in Fig. 9 we highlight how different methods create robustness through different dynamics in the loss and feature utilization landscape, as discussed further in suppl. material, where we also share further details and results with other models and datasets, as well as more background on FL and SAM.

Does LR have similar effects on generalization in standard classification tasks? We next investigate whether the robustness against SCs afforded by high LRs could help explain their success in generalization under standard classification tasks [3, 43, 51]. A positive answer to this question would imply that in realistic machine learning settings, even a seemingly IID generalization task can involve OOD generalization subtasks due to implicit and/or rare SCs in the training distribution (see e.g. [63]). We start our analyses with CIFAR-10 and CIFAR-100, and then move on to ImageNet-1k.

To qualitatively investigate whether the generalization advantage of large LRs in naturalistic classification tasks pertain to SCs, under a ResNet18 model, we examine input attributions for samples most likely to be predicted correctly by high LR models compared to low LRs (see suppl. material for details), presented in Fig. 8. Our results show that high LR models are likelier to focus on core vs. spurious features not only in the semi-synthetic MNIST-CIFAR dataset, but also in the CIFAR-10 and CIFAR-100 datasets as well. In both these datasets, low LR models focus much more on backgrounds and/or the color/texture of the foreground object, whereas the high LR models are likelier to focus on the object contours. The results strongly suggest that high LRs might be obtaining increased generalization through robustness to hidden SCs in the training data. We thus proceed to test this hypothesis quantitatively.

Quantitatively measuring SCs in naturalistic datasets remains an open problem [83]. Nevertheless, here we develop two complementary metrics to assess the effects of LR on spurious feature utilization. The first, *attribution entropy* relies on computing the entropy of a normalized input attribution map since exploiting background information or object textures/colors would lead to more dispersed attribution maps as opposed focusing on contours of an object. We complement this relatively simple metric with a more targeted one in *background attribution percentage*: Utilizing the DeepLabV3 segmentation algorithm to extract background maps [7], we compute the percentage of attributions that fall within vs. outside these image background segments. We present the results in Fig. 10. The results not only replicate the benevolent effects of large LRs seen in previous datasets, they also show that higher LR models are less likely to use background/texture information.

ImageNet-1k. We lastly examine the effects of large LRs on models trained on ImageNet-1k dataset. The results presented in Fig. 11 clearly show that the previously observed effects of LR extend to larger datasets, furthering the implications of our findings to realistic scenarios.

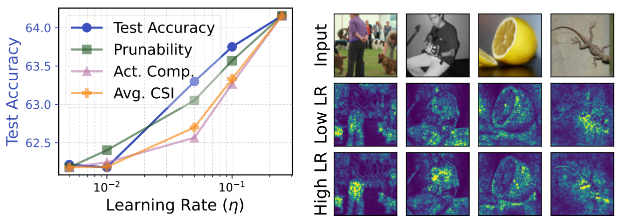


Figure 11. Effects of LR with ImageNet-1k dataset and ResNet18 model on model statistics and attributions.

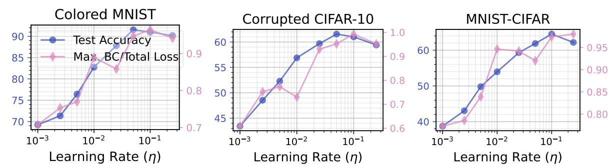


Figure 12. Unbiased test accuracy strongly correlates with maximum bias-conflicting loss ratio during training.

6. On the mechanism of large LRs

While plausible mechanisms through which large LRs create compressibility have been proposed [24], those through which large LRs can create robustness against SCs are much less explored, which we now investigate. Previous work indicates that simple spurious features are learned earlier than more complex core features [57, 60], regardless of LR [82, Theorem 4.1]: this produces a phase in learning where bias-aligned (BA) samples are predicted correctly, while bias-conflicting (BC) samples are misclassified. We find that at this stage, another effect of large LRs, namely fast *norm growth* of model parameters and logits, become important[4]. This is because with cross-entropy (CE) loss, up-scaled logits under high LRs lead to *confident mispredictions of BC samples*. Due to the nonlinearty of CE, the mini-batch loss (and thus the gradient) is then increasingly dominated by mispredicted BC samples, corresponding to an implicit *reweighting of the dataset in favor of BC samples*. We now formalize this statement.

Proposition 1 *Let $f_{\mathbf{w}}$ be a predictor in a binary classification task, trained under the cross-entropy loss ℓ. Let $\Omega = \{\mathbf{z}_1 \dots \mathbf{z}_{|\Omega|}\}$ be a training mini-batch such that $(\mathbf{x}_i, y_i, b_i) \overset{i.i.d.}{\sim} \mu_Z^{\text{train}}$. Let $\Omega_{\text{bc}} := \{\mathbf{z}_i \in \Omega | y_i \neq b_i\}$ and $\Omega_{\text{ba}} := \{\mathbf{z}_i \in \Omega | y_i = b_i\}$ represent the bias-conflicting and bias-aligned subsets of Ω. If $f_{\mathbf{w}}$ predicts according to the spurious decision rule,* i.e. $b = \arg\max_j f_{\mathbf{w}}(\mathbf{x})[j]$, *then for some $\alpha > 0$, as the logit-scaling factor $k \to \infty$*

$$\frac{\sum_{\mathbf{x},y \in \Omega_{\text{bc}}} \ell(y, k f_{\mathbf{w}}(\mathbf{x}))}{\sum_{\mathbf{x},y \in \Omega_{\text{ba}}} \ell(y, k f_{\mathbf{w}}(\mathbf{x}))} = \mathcal{O}(k e^{\alpha k}). \tag{1}$$

The proof is deferred to the suppl. material, which also presents another proposition, showing how this can result in an increase in the gradient norms to the subnetworks that utilize spurious vs. core features. These imply that large LRs generate strong gradients that discourage reliance on spurious features, whereas small LRs lack this pressure, allowing models to retain spurious subnetworks while memorizing BC samples. Through experiments with ResNet18, Fig. 12 demonstrate that unbiased test accuracy *very strongly* correlates with the maximum loss ratio (*i.e.* BC loss/total minibatch loss) encountered during training across datasets, supporting our proposed mechanism (see suppl. material for similar results with additional datasets and models). Fig. 13 visualizes training dynamics under a large vs. small LR ($\eta = 0.001$ vs 0.1): early focus on spurious features are weaned off under large LRs, as BC samples dominate the loss due to confident mispredictions, both in Colored MNIST dataset *and* CIFAR-10 datasets.

[4] While norm growth under large LRs is well-established by previous research [49, 52, 66], its effects are underexplored in robust generalization.

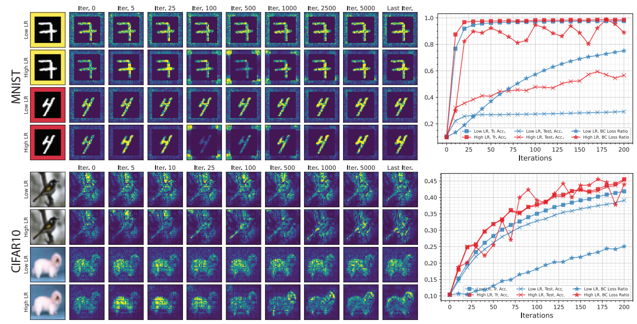


Figure 13. Large LRs lead to losses from bias-conflicting samples to dominate the gradient, resulting in increased utilization of core features. (Left) Attribution maps (right) model statistics for Colored MNIST and CIFAR-10 through training.

We can interpret these large, disruptive updates to the network under the outsized influence of bias-conflicting examples as a recurring *lottery ticket* type scenario [20], where spurious subnetworks are effectively reset by large gradient updates whenever they rely too much on spurious feature and lead to confident mispredictions, similar to the "catapult mechanism", proposed by [40].

7. Conclusion

Robustness to SCs and compressibility are crucial requirements for advancing safe, trustworthy, and resource-efficient deep learning. In the absence of strong theoretical guarantees, our investigation demonstrates the efficacy of LR as a significant inductive bias for addressing these concerns simultaneously. We show that high LRs jointly obtain robustness to spurious correlations, network compressibility, and favorable representation properties.

Limitations and future work. Our work is far from an end-to-end, convergent explanation involving optimization dynamics, parameter loss landscapes, and representations. Future work should also study how multiple/hierarchical SCs interact with training dynamics, and design explicit training procedures for more robust, efficient models.

Acknowledgments

M. Barsbey was supported by the EPSRC Project GNOMON (EP/X011364/1). L. Prieto was supported by the UKRI Centre for Doctoral Training in Safe and Trusted AI [EP/S0233356/1]. S. Zafeiriou and part of the research was funded by the EPSRC Fellowship DEFORM (EP/S010203/1) and EPSRC Project GNOMON (EP/X011364/1). T. Birdal acknowledges support from the Engineering and Physical Sciences Research Council [grant EP/X011364/1]. T. Birdal was supported by a UKRI Future Leaders Fellowship [grant MR/Y018818/1] as well as a Royal Society Research Grant [RG/R1/241402].

References

[1] Maksym Andriushchenko, Aditya Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with large step sizes learns sparse features. In *Proceedings of the 40th International Conference on Machine Learning*, pages 903–925, Honolulu, Hawaii, USA, 2023. JMLR.org. 2, 1, 3

[2] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger Generalization Bounds for Deep Nets via a Compression Approach. In *International Conference on Learning Representations*, 2018. 2, 1

[3] Melih Barsbey, Milad Sefidgaran, Murat A Erdogdu, Gaël Richard, and Umut Şimşekli. Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks. In *Advances in Neural Information Processing Systems*, 2021. 2, 7, 1

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in Terra Incognita. In *European Conference on Computer Vision*, pages 472–489, 2018. 2, 1

[5] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Şimşekli. Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, 2021. 2, 1

[6] Sebastien Bubeck and Mark Sellke. A Universal Law of Robustness via Isoperimetry. In *Advances in Neural Information Processing Systems*, 2021. 1

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv:1706.05587*, 2017. 7

[8] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *International Conference on Learning Representations*, 2021. 2, 1

[9] Jude Coleman. AI's Climate Impact Goes beyond Its Emissions. https://www.scientificamerican.com/article/ais-climate-impact-goes-beyond-its-emissions/, 2023. 1

[10] Josh Constine and Veronica Mercado. The AI compute shortage explained by Nvidia, Crusoe, & MosaicML | AI Venture Capital. https://www.signalfire.com/blog/ai-compute-shortage, 2023. 1

[11] 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, 2009. 5

[12] Benoit Dherin, Michael Munn, Mihaela Rosca, and David Barrett. Why neural networks find simple solutions: The many regularizers of geometric complexity. *Advances in Neural Information Processing Systems*, 35:2333–2349, 2022. 1

[13] Enmao Diao, Ganghua Wang, Jiawei Zhang, Yuhong Yang, Jie Ding, and Vahid Tarokh. Pruning Deep Neural Networks from a Sparsity Perspective. In *The Eleventh International Conference on Learning Representations*, 2023. 6

[14] James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A Winning Hand: Compressing Deep Networks Can Improve Out-of-Distribution Robustness. In *Advances in Neural Information Processing Systems*, 2021. 2, 1

[15] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp Minima Can Generalize For Deep Nets. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017. 2

[16] Thu Dinh, Bao Wang, Andrea Bertozzi, Stanley Osher, and Jack Xin. Sparsity Meets Robustness: Channel Pruning for the Feynman-Kac Formalism Principled Robust Deep Neural Nets. In *Machine Learning, Optimization, and Data Science*, pages 362–381, Cham, 2020. Springer International Publishing. 1

[17] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut Learning of Large Language Models in Natural Language Understanding. *Commun. ACM*, 67(1):110–120, 2023. 2

[18] Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. Robustness Challenges in Model Distillation and Pruning for Natural Language Understanding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1766–1778, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. 2, 1

[19] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*, 2021. 7

[20] Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*, 2019. 8

[21] Brian Fung. The big bottleneck for AI: A shortage of powerful chips | CNN Business. https://www.cnn.com/2023/08/06/tech/ai-chips-supply-chain/index.html, 2023. 1

[22] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2, 4, 1

[23] Rémi Gribonval, Volkan Cevher, and Mike E Davies. Compressible distributions for high-dimensional statistics. *IEEE Transactions on Information Theory*, 58(8):5016–5034, 2012. 3

[24] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu.

The Heavy-Tail Phenomenon in SGD. In *International Conference on Machine Learning*, 2021. 8

[25] Katherine Haan. Most Common Way Consumers Plan to Use Artificial Intelligence. https://datawrapper.dwcdn.net/R4rT3/4/, 2024. 1

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[27] Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. In *First Conference on Language Modeling*, 2024. 6

[28] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022. 2, 1

[29] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho*, and Krzysztof Geras*. The Break-Even Point on Optimization Trajectories of Deep Neural Networks. In *International Conference on Learning Representations*, 2020. 2, 1

[30] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *arXiv:1609.04836*, 2017. 2, 1, 3

[31] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[32] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In *The Eleventh International Conference on Learning Representations*, 2022. 4, 6, 5

[33] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *arXiv:2009.07896*, 2020. 7

[34] Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why Do Better Loss Functions Lead to Less Transferable Features? In *Advances in Neural Information Processing Systems*, 2021. 4

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 5

[36] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks. In *International Conference on Machine Learning*, 2021. 7

[37] Anisio Lacerda, Daniel Ayala, Francisco Malaguth, and Fabio Kanadani. An Empirical Analysis of Vision Transformers Robustness to Spurious Correlations in Health Data. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023. 2

[38] Matthew L. Leavitt and Ari S. Morcos. Selectivity considered harmful: Evaluating the causal impact of class selectivity in DNNs. In *International Conference on Learning Representations*, 2021. 4

[39] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[40] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism. *arXiv:2003.02218*, 2020. 2, 8, 1

[41] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning Filters for Efficient ConvNets. In *International Conference on Learning Representations*, 2017. 3

[42] Xi-Lin Li. Preconditioned Stochastic Gradient Descent. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1454–1466, 2018. 6

[43] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks. In *Neural Information Processing Systems*, 2019. 2, 7, 1

[44] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 7

[45] Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. In *International Conference on Machine Learning*, 2021. 1, 2, 3

[46] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 5

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv:2103.14030*, 2021. 6

[48] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*, 2017. 6

[49] Clare Lyle, Zeyu Zheng, Khimya Khetarpal, James Martens, Hado van Hasselt, Razvan Pascanu, and Will Dabney. Normalization and effective learning rates in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024. 8

[50] Hannes Mandler and Bernhard Weigand. A review and benchmark of feature importance methods for neural networks. *ACM Comput. Surv.*, 56(12):318:1–318:30, 2024. 6

[51] Charles H. Martin and Michael W. Mahoney. Traditional and Heavy-Tailed Self Regularization in Neural Network Models. *arXiv:1901.08276*, 2019. 7

[52] William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. 8

[53] Amirkeivan Mohtashami, Martin Jaggi, and Sebastian Stich. Special Properties of Gradient Descent with Large Learning Rates. In *Advances in Neural Information Processing Systems*, 2023. 2, 1

[54] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2022. 2, 3, 1

[55] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep Double Descent: Where Bigger Models and More Data Hurt. In *Eighth International Conference on Learning Representations*, 2020. 2

[56] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: Training Debiased Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*, 2020. 3, 4, 5, 1, 2

[57] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. In *Advances in Neural Information Processing Systems*, 2021. 2, 8, 1

[58] Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, and Guillaume Lajoie. Multi-scale Feature Learning Dynamics: Insights for Double Descent. In *International Conference on Machine Learning*, 2022. 1

[59] Aahlad Manas Puli, Lily H. Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame Dataset Shift! Shortcut Learning due to Gradients and Cross Entropy. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 4, 1

[60] Guanwen Qiu, Da Kuang, and Surbhi Goel. Complexity Matters: Feature Learning in the Presence of Spurious Correlations. In *International Conference on Machine Learning*, 2024. 3, 4, 8, 5

[61] Anant Raj, Melih Barsbey, Mert Gurbuzbalaban, Lingjiong Zhu, and Umut Şimşekli. Algorithmic Stability of Heavy-Tailed Stochastic Gradient Descent on Least Squares. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, pages 1292–1342, 2023. 2

[62] Omkar Ranadive, Nikhil Thakurdesai, Ari S. Morcos, Matthew L. Leavitt, and Stephane Deny. On the special role of class-selective neurons in early training. *Transactions on Machine Learning Research*, 2023. 4

[63] Elan Rosenfeld and Andrej Risteski. Outliers with Opposing Signals Have an Outsized Effect on Neural Network Optimization. In *International Conference on Learning Representations*, 2024. 2, 7, 1, 3

[64] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *International Conference on Learning Representations*, 2020. 2, 5, 1

[65] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8346–8356. JMLR.org, 2020. 2, 4, 1

[66] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, 2016. 8

[67] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. In *Advances in Neural Information Processing Systems*, pages 9573–9585, 2020. 2, 5, 1

[68] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 3145–3153, Sydney, NSW, Australia, 2017. JMLR.org. 7

[69] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*, 2015. 5, 3

[70] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 3319–3328, Sydney, NSW, Australia, 2017. JMLR.org. 6

[71] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning - Volume 70*, 2017. 4

[72] Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura. Spectral Pruning: Compressing Deep Neural Networks via Spectral Analysis and its Generalization Error. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2839–2846. International Joint Conferences on Artificial Intelligence Organization, 2020. 1

[73] Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: Unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020. 1

[74] Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In *International Conference on Machine Learning*, pages 34330–34343. PMLR, 2023. 2, 1

[75] Vishaal Udandarao, Karsten Roth, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier Hénaff, Samuel Albanie, Zeynep Akata, and Matthias Bethge. A Practitioner's Guide to Real-World Continual Multimodal Pretraining. *Advances in Neural Information Processing Systems*, 37:133801–133845, 2024. 2

[76] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[77] Luyu Wang, Gavin Weiguang Ding, Ruitong Huang, Yanshuai Cao, and Yik Chau Lui. Adversarial Robustness of Pruned Neural Networks. 2018. 1

[78] Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing Sharpness along GD Trajectory: Progressive Sharpening and Edge of Stability. In *Advances in Neural Information Processing Systems*, pages 9983–9994, 2022. 2, 1

[79] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A Fine-Grained Analysis on Distribution Shift. In *International Conference on Learning Representations*, 2021. 1, 3

[80] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. 1

[81] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022. 6

[82] Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying Spurious Biases Early in Training through the Lens of Simplicity Bias. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 2953–2961. PMLR, 2024. 8

[83] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious Correlations in Machine Learning: A Survey. *arXiv:2402.12715*, 2024. 5, 7

[84] Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. *arXiv:1605.07146*, 2017. 5

[85] Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: Spikes in the training loss and their impact on generalization through feature learning. In *International Conference on Machine Learning*, 2024. 1