

The Effect of Model Size on Worst-Group Generalization

Alan Pham

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-138

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-138.html>

May 18, 2022



Copyright © 2022, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

The Effect of Model Size on Worst-Group Generalization

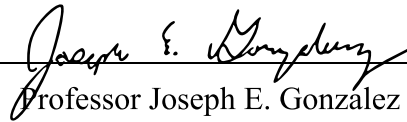
by Alan Pham

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences,
University of California at Berkeley, in partial satisfaction of the requirements for the
degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:



Professor Joseph E. González
Research Advisor

5/12/2022

(Date)

* * * * *



Professor Jacob Steinhardt
Second Reader

05/12/2022

(Date)

The Effect of Model Size on Worst-Group Generalization

by

Alan Pham

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Joseph E. Gonzalez, Chair

Professor Jacob Steinhardt

Spring 2022

Abstract

The Effect of Model Size on Worst-Group Generalization

by

Alan Pham

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Joseph E. Gonzalez, Chair

Overparameterization is shown to result in poor test accuracy on rare subgroups under a variety of settings where subgroup information is known. To gain a more complete picture, we consider the case where subgroup information is unknown. We investigate the effect of model size on worst-group generalization under empirical risk minimization (ERM) across a wide range of settings, varying: 1) architectures (ResNet, VGG, or BERT), 2) domains (vision or natural language processing), 3) model size (width or depth), and 4) initialization (with pre-trained or random weights). Our systematic evaluation reveals that increasing model size does not hurt, and may help, worst-group test performance under ERM across all setups. In particular, increasing pre-trained model size consistently improves performance on Waterbirds and MultiNLI. We advise practitioners to use larger pre-trained models when subgroup labels are unknown.

Contents

Contents	i
List of Figures	ii
List of Tables	iv
1 The Effect of Model Size on Worst-Group Generalization	1
1.1 Introduction	2
1.2 Related Work	4
1.3 Problem Setting	5
1.4 Experimental Setup	6
1.5 Experimental Results	10
1.6 Conclusion	15
Bibliography	16

List of Figures

- 1.1 Models trained to convergence, e.g., until average training accuracy is close or equal to 100%. Hyperparameters remain the same within each experiment series. In each of the four graphs above, we compare the error of pre-trained and randomly initialized models of the same architecture. Pre-trained models perform better than models trained from scratch and increasing model sizes for both types of models does not hurt the worst-group error. **Top row:** Depth-varying results. **Bottom row:** Width-varying results. **Columns:** From left to right, ResNet on Waterbirds, ResNet on CelebA Blond / Male, ResNet on CelebA Lipstick / Earring, BERT on MultiNLI. 10
- 1.2 The top row shows the pre-trained ResNet models of varying depth. The bottom row shows the pre-trained VGG models of varying depth. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For the two CelebA datasets, model depth has a negligible effect on the worst-group error whereas on the Waterbirds dataset, the increasing the model size decreases the worst-group error. 11
- 1.4 Depth of randomly initialized ResNet models is varied, increasing in depth from left to right. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For all of the datasets, model depth has a negligible effect on the worst-group error. 12
- 1.3 The top row shows the pre-trained ResNet18 models of varying widths. The bottom row shows the pre-trained MobileNet models of varying width. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For the two CelebA datasets, model depth has a negligible effect on the worst-group error whereas on the Waterbirds dataset, the increasing the model size decreases the worst-group error. 12

- 1.5 Width of randomly initialized ResNet models is varied, increasing in width from left to right. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For all of the datasets, model depth has a negligible effect on the worst-group error. 13
- 1.6 **Top Row:** Depth and width of pre-trained BERT models are varied, increasing in size from left to right. **Bottom Row:** Depth and width of randomly initialized BERT models are varied, increasing in size from left to right. Increasing pre-trained model size reduces worst-group error, while on randomly initialized models, model size has negligible effect. 13
- 1.7 Graphs displaying how the error changes as we vary the width or number of features. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For CelebA Blond / Male and Waterbirds, worst-group error improves with the greater model size whereas worst-group error on the CelebA Lipstick / Earring seems to stay at about the same value as model size increases. 14

List of Tables

1.1	Number of training examples in each dataset	8
1.2	The hyperparameters used to train the models to convergence on each dataset. 9	
1.3	Comparison of group DRO summarized from Sagawa et al. [26] with our own ERM results.	9

Acknowledgments

Thank you so much to everyone who has supported me in my educational journey. This work was a joint effort with students Eunice Chan, Vikranth Srivatsa, Dhruba Ghosh; postdoctoral researcher Yaoqing Yang; graduate students Yaodong Yu, Ruiqi Zhong; and advisors Professor Joseph Gonzalez and Professor Jacob Steinhardt [24]. Thank you to my advisor Professor Joseph Gonzalez for his valuable advice and support; to Professor Jacob Steinhardt and all my other professors for their support over the years; to Yaoqing Yang, Yaodong Yu, Moustafa AbdelBaky, Daniel Rothchild and others for their mentorship; to Pang Wei Koh, Aditi Raghunathan, and Erik Jones for their valuable feedback and comments during preparation of this work; to my friends for their support; and to my family for their encouragement and love.

Chapter 1

The Effect of Model Size on Worst-Group Generalization

1.1 Introduction

Recent work shows that overparameterized models display stronger generalization performance than smaller models, even on small datasets, suggesting that larger models overfit less [2]. Despite this trend, many works find or assume that overparameterization can hurt test accuracy on certain subgroups of the data [3, 7, 20, 26, 36]. This problem is exacerbated when tuning a model with unknown subgroup labels since the performance gap between average accuracy and worst-group accuracy—the subgroup with poorest accuracy—can be large. The difference between the average test accuracy and worst-group test accuracy ranges from 40 to 60% in the empirical risk minimization (ERM) setup from Sagawa et al. [25] on CelebA and Waterbirds. In addition to the resulting fairness concerns, real-world deployments might encounter distribution shifts that upweight rare subgroups, causing the average accuracy to suffer.

Low worst-group accuracy often occurs when subgroups are associated with spurious features. For example, consider the task of classifying images of landbirds and waterbirds: although waterbirds are more likely to appear on a water background, the image background (the spurious feature) has no direct causal relationship with the species of bird. These spurious features harm performance on “rare” subgroups (e.g. waterbirds on land background) where the cue contradicts the true label. This can lead to high-stakes real-world problems: thoracic pathology detection models were found to rely on the presence of chest drains (a treatment device) to detect pneumothorax, which led to poor performance on the clinically relevant rare subgroup of untreated patients with pneumothorax [23].

We systematically investigate the effect of model size on the performance of the model on the rare subgroups. Prior works show that increasing model size can hurt worst-group performance. In Sagawa et al. [25], this trend was shown for models trained with the reweighted objective that upweights minority groups. Furthermore, Sagawa et al. [25] finds that, trained with naive ERM, models have poor worst-group error regardless of model size. In our work, we provide complementary findings as we explore the trend in naive ERM models in more comprehensive and informative experiments. We experiment on a wide range of datasets with spurious correlations, which include: Waterbirds (a dataset of birds on land and water) [26], CelebA (a face dataset we use for two tasks: Lipstick / Earring and Blond / Male) [18], and MultiNLI (a language dataset) [34] datasets (more details and conventions listed in Appendix 1.4). We also conduct experiments across multiple model architectures, varying model depth or width and initializing with pre-trained or random weights.

We find that, under ERM, increasing model size does not hurt and sometimes helps worst-group test accuracy across all settings considered in this paper (Section 1.5). In particular, larger pre-trained models are often less susceptible to spurious correlation. For example, in the Waterbirds dataset, compared to pre-trained ResNet18, pre-trained ResNet152 decreases the worst-group error by approximately 18.39% (Figure 1.2).

Interestingly, pre-trained models actually achieve *better* worst-group accuracy as model size increases, while models trained from scratch merely do not get worse.

We summarize our contributions as follows:

- We empirically show that, under ERM, larger model sizes either help or do not hurt worst-group test accuracy across a wide range of settings. Specifically, we look across vision and language datasets and common model architectures.
- We explore the trend over increasing numbers of parameters by separately looking at the effects widening models and deepening models. We also look at the effects of pre-training the model on a different dataset compared to randomly initialized models.
- We find that larger pre-trained models consistently improve worst-case group accuracies on two widely used datasets: Waterbirds and MultiNLI.

1.2 Related Work

Generalization with Overparameterized Models. Many widely-used models, including the ResNet, VGG batchnorm (BN), and BERT series of architectures, show a monotonic increase in average test accuracy on common tasks with increasing model size. Overparameterization, increasing model size beyond memorizing the training set, is also found to improve robustness to adversarial samples and distributional shifts [11]. This has been attributed to the “double descent”, where increasing model size results in worse performance, then better model performance, past some interpolation threshold [1, 21]. Bornschein, Visin, and Osindero [2] discusses this in the context of small data, while Yang et al. [35] provides an explanation via the bias-variance trade-off curve. Sagawa et al. [25] studies overparameterization and worst-group accuracy, finding that larger models often fail to generalize to rare groups in the test distribution. This is attributed to the presence of spurious correlations in the training data.

Distributional Shift and Spurious Correlation. We examine the phenomenon of distributional shift, where training and test set distributions differ [15]. Pre-trained language models are more robust against spurious correlation, though this is dependent on the number of negative samples (i.e., samples that for which the spurious correlation does not correctly predict the label) [29]. On the other hand, Niven and Kao [22] finds that language models heavily rely on spurious correlations. For instance, on the argument reasoning comprehension task, even when logically necessary parts of the input are obscured, BERT Large is still reliably able to perform well. Meanwhile, vision models suffer from spurious correlations due to scene biases such as co-occurrence of objects with other objects, backgrounds, or textures [38, 12].

Mitigating Spurious Correlation. Prior works have explored data augmentation to combat dependence on spurious correlations. In NLP, one approach is counterfactual augmentation: modifying the input sentences by a minimal amount to change the target label [13]. This reduces the strength of spurious correlations by introducing variation in core features (with a causal relation to the label) while controlling for non-causal variables. In computer vision, Goel et al. [5] suggests using CycleGAN to generate negative examples. Liu et al. [17] suggests a two-stage training regime that upweights misclassified points. Wang et al. [32] introduces a causal attention module that performs unsupervised annotations to mitigate spurious correlation.

Alternatively, worst-group test error can be reduced through the use of distributionally robust optimization (DRO) to guide training [38, 8]. One approach is group DRO (GDRO), an instance of DRO that minimizes the worst-group expected loss [26, Eqn.(4)]. This substantially improves generalization when coupled with heavy regularization, but requires prior labeling of subgroups in the training set.

1.3 Problem Setting

We adopt the formulation presented in Sagawa et al. [25]. Each sample consists of an input $x \in \mathcal{X}$, a target label $y \in \mathcal{Y}$, and a spurious attribute $a \in \mathcal{A}$. We categorize samples into groups $g = (y, a) \in \mathcal{G} = \mathcal{Y} \times \mathcal{A}$. The spurious attribute a is correlated with the label y , but has no causal relationship. The problem domains we consider are all classification tasks, where $|\mathcal{Y}| = 2$ or 3 , and are confounded by a binary spurious attribute ($|\mathcal{A}| = 2$). For example, in the Waterbirds dataset, we might have an image x of a waterbird on a land background: the label y is the target class “waterbird”, the spurious feature a is “land background”, and the group g is (waterbird, land background).

Our work focuses on the effect of model size on worst-group test error. We train using empirical risk minimization (ERM) [31], finding the model parameters θ that empirically minimize the average training loss:

$$\mathcal{R}_{\text{ERM}}(\theta) = \mathbb{E}_{(x,y,g)}[\ell(\theta, (x, y))]. \quad (1.1)$$

We use cross-entropy loss and choose the weight decay, learning rate, and training epoch so that the models are trained to convergence. Further details on the training procedure are in Appendix 1.4.

Metrics. Given a model $h : \mathcal{X} \rightarrow \mathcal{Y}$, we define the error on a group $g \in \mathcal{G}$ as

$$\varepsilon_g := \mathbb{E}_{x,y|g}[\mathbf{1}(h(x) \neq y)]. \quad (1.2)$$

Throughout the paper, we compute the group error averaged over the last 10 epochs of training in order to reduce noise and smooth out the results.

We consider two metrics: the average error as well as the worst-group error, defined as:

$$\varepsilon := \sum_{g \in \mathcal{G}} w_g \cdot \varepsilon_g \quad \text{and} \quad \varepsilon_{\text{wg}} := \max_{g \in \mathcal{G}} \varepsilon_g, \quad (1.3)$$

where w_g is the weight equal to the group’s proportion in the training data.

Across our experiments on increasing model depths and widths for the ResNet, VGG BN, MobileNet, and BERT architectures, we report average and worst-group error on the training and test datasets.

1.4 Experimental Setup

We compare the trends of the average and worst-group train and test performance of different model architectures by varying the model size in terms of depth and width. For CV, we use ResNet [9], VGG (batchnorm) [28], and MobileNet [27]. For the NLP domain, we use BERT [4]. Compared to prior work, we use a wide range of commonly-used models including those run on edge devices to observe the trends in realistic settings. We also make sure that we follow the experiment settings in prior works. For example, the setup over the CelebA datasets follows the setup of Nakkiran et al. [21], varying the width of a ResNet10 model [10]. The setup over the Waterbirds dataset follows the setup in Mei and Montanari [19], training an unregularized logistic regression model over a variable number of projections of the feature representation of the input in a pre-trained ResNet18 model.

Datasets

We look at the trends on four tasks: Waterbirds, CelebA Blond / Male, CelebA Lipstick / Earring, and MultiNLI. The first three are CV tasks whereas MultiNLI is a NLP task.

Waterbirds. Waterbirds is a synthetic dataset constructed in [26] by cropping out bird photographs from the Caltech-UCSD Birds-200-2011 (CUB) dataset [33] and placing them on top of image backgrounds from the Places dataset [37]. In Waterbirds, we classify two types of birds: birds that primarily live on land (landbirds) and birds that live on water (waterbirds). These classes are spuriously correlated with the background type: land background or water background.

Most landbirds are photographed on land and waterbirds on water. Therefore, there are four groups of varying sizes in this dataset. Two large groups of common pairings: landbird on land background, and landbird on water background, and two small groups of less common pairings: waterbird on land background, and waterbird on water background. The background acts as a spurious factor; models typically associate water backgrounds with waterbirds and vice versa. As a result, models tend to fail to generalize to rare groups.

CelebA [18] is a large scale multi-feature face dataset with varying backgrounds and poses. Using the CelebA dataset, we can construct spurious correlation datasets by selecting specific features from the multi-feature dataset. CelebA Blond / Male and CelebA Lipstick / Earring are two such examples of spurious correlation datasets formed from CelebA.

CelebA Blond / Male. In CelebA Blond / Male, the model classifies images as either containing blond or dark hair. The model classifies images as either containing blond hair or not. The spurious correlation is whether or not the subject is male.

CelebA Lipstick / Earring. In the CelebA Lipstick / Earring task, the model classifies images as either containing lipstick or no lipstick [14]. The spurious correlation is the presence of earrings, which is highly correlated with the presence of lipstick.

MultiNLI. MultiNLI is a natural language inference dataset introduced in [34]. The NLI task consists of predicting how a sentence A logically relates to another sentence B.

While the three labels (entailment, contradiction, neutral) are represented equally in the dataset, [6] discovers an annotation artifact: negation words (“nobody”, “no”, “never”, and “nothing”), when present in sentence B, are far more likely to correspond to a contradiction than entailment. Thus, the labels are spuriously correlated with the presence of negation words.

Table 1.4 in the appendix describes the number of samples in each group.

Models

We refer to “pre-trained” models as those where we finetuned models which were pre-trained (say on ImageNet for CV tasks), or “trained from scratch” models where our initial checkpoint begins with randomly initialized weights.

VGG (batchnorm). VGG batchnorm (BN) models are large CNN, which extend on AlexNet using multiple 3x3 sized filters. It was one the of the top performing models for the ImageNet image localization task. We train and test pretrained VGG BN 11, VGG BN 13, VGG BN 16, and VGG BN 18 models.

ResNet. ResNet is a state of the art deep convolutional model that also perform very well on image classification on ImageNet. It was created to minimize the vanishing gradient by adding a residual block that adds weights from previous layers. For the depths, we train and test pretrained ResNet18, ResNet34, ResNet101, and ResNet152 models. For the widths, we train and test pretrained ResNet18 WD4, ResNet18 WD2, ResNet18 W3D4, and ResNet18 models.

MobileNet V2 width reductions. Different width multipliers applied all layers (but the last convolutional layer) of the MobileNetv2 architecture, pretrained on ImageNet [27]. We use architectures and pretrained checkpoints from a Github reproduction from [16].

BERT. We use the BERT architecture for the MultiNLI task, varying the width and depth as in [30], which showed effectiveness of BERT even with non-standard depths and widths.

Training Procedure

The CV models (ResNet, VGG, MobileNet) were trained with Nvidia GPUs. The BERT models were pretrained and fine-tuned on a TPUv3 through the Google Cloud Platform. For the CV models, we use a batch size of 128, a stochastic gradient descent (SGD) optimizer with a momentum of 0.9, and a step scheduler for the learning rate. For the BERT models, we use a batch size of 64, the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a linear LR warmup for 10% of the training epochs followed by linear LR decay to zero. A list of the hyperparameters used can be found in Table 1.2.

Resampling

We resample each dataset to verify the consistency of the trends. 95% confidence interval error bars are included in the graphs for context. Resampling is done by shuffling train, validation, and test data while keeping the proportions of the 4 subgroups in each set. The models are then trained to convergence and we graph the training and validation errors of the converged models. Within each set of model size experiments (i.e., set of graphs), we trained all the models using the same hyperparameters.

Dataset	Worst-Group	Total #	Worst-Group #
Waterbirds	Waterbird, Land Background	4795	56
CelebA Blond / Male	Blond, Male	162770	1387
CelebA Lipstick / Earring	No Lipstick, Earring	162770	4516
MultiNLI	Entailment, Negation	205357	1483

Table 1.1: Number of training examples in each dataset

Dataset	Initial Weight	Architecture	Epochs	LR	LR Step	WD
Waterbirds	Pretrained	ResNet	100	5e-4	30	1e-4
		ResNet Width	100	0.001	30	1e-4
		MobileNet	100	0.01	30	1e-4
		VGG BN	100	5e-4	30	1e-4
	Random	ResNet	100	0.01	30	1e-4
		ResNet Width	100	0.01	30	1e-4
CelebA Blond / Male	Pretrained	ResNet	100	5e-4	30	1e-4
		ResNet Width	100	0.001	30	1e-4
		MobileNet	200	0.05	50	3e-5
		VGG BN	100	5e-4	30	1e-4
	Random	ResNet	100	0.01	30	1e-4
		ResNet Width	100	0.01	30	1e-4
CelebA Lipstick / Earring	Pretrained	ResNet	100	5e-4	30	1e-4
		ResNet Width	100	0.001	30	1e-4
		MobileNet	200	0.06	50	3e-5
		VGG BN	100	5e-4	30	1e-4
	Random	ResNet	100	0.01	30	1e-4
		ResNet Width	100	0.01	30	1e-4
MultiNLI	Pretrained	BERT	20	5e-5	1	0.01
	Random	BERT	20	5e-5	1	0.01

Table 1.2: The hyperparameters used to train the models to convergence on each dataset.

Dataset	Worst-Group DRO Acc.	Worst-Group ERM Acc.
Waterbirds	84.6	39.47
CelebA Blond/Male	88.3	58.41
MultiNLI	77.7	67.5

Table 1.3: Comparison of group DRO summarized from Sagawa et al. [26] with our own ERM results.

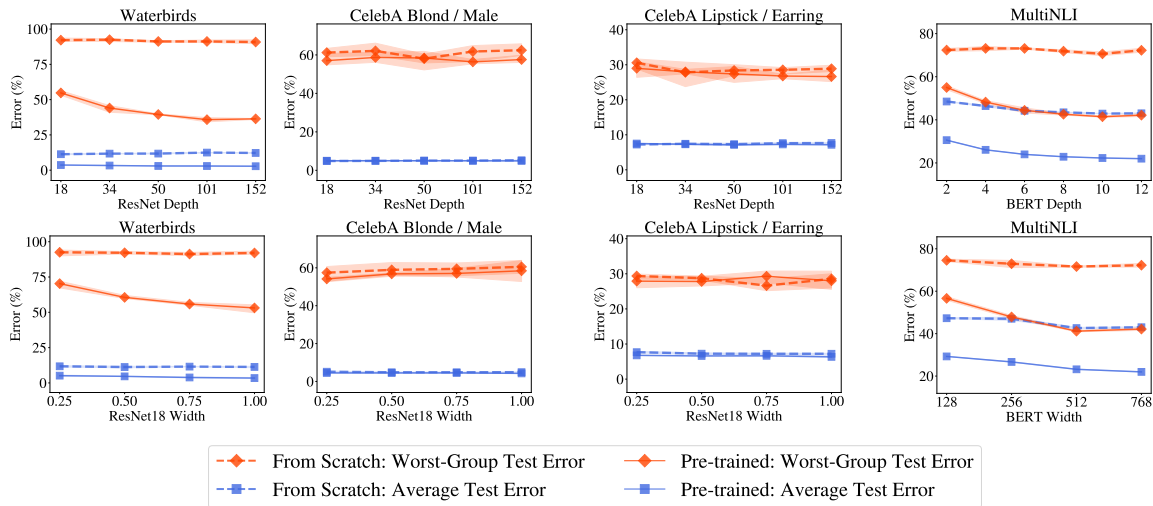


Figure 1.1: Models trained to convergence, e.g., until average training accuracy is close or equal to 100%. Hyperparameters remain the same within each experiment series. In each of the four graphs above, we compare the error of pre-trained and randomly initialized models of the same architecture. Pre-trained models perform better than models trained from scratch and increasing model sizes for both types of models does not hurt the worst-group error. **Top row:** Depth-varying results. **Bottom row:** Width-varying results. **Columns:** From left to right, ResNet on Waterbirds, ResNet on CelebA Blond / Male, ResNet on CelebA Lipstick / Earring, BERT on MultiNLI.

1.5 Experimental Results

We are interested in understanding how worst-group test accuracy changes with model size. To investigate the trends systematically, we study the effect of varying width and depth, under pre-trained and randomly initialized models. We perform four sets of experiments, varying the depth and width of pre-trained and randomly initialized models. The summarized results of the experiments on the Waterbirds and MultiNLI datasets are presented in Figure 1.1.

We summarize our depth varying for pre-trained models results in figure 1.2, width varying for pre-trained models in figure 1.3, depth varying for randomly initialized models in figure 1.4, and width varying for randomly initialized models in figure 1.5. Pre-trained models on the Waterbirds dataset show a distinct decrease in error with model size while for other experiments, error stays roughly the same for all model sizes.

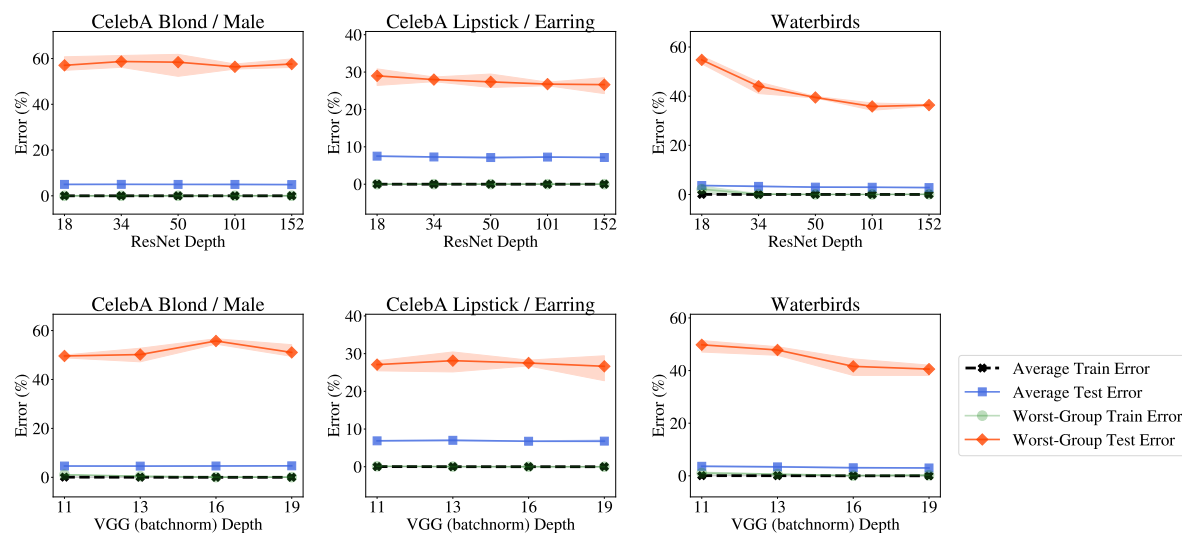


Figure 1.2: The top row shows the pre-trained ResNet models of varying depth. The bottom row shows the pre-trained VGG models of varying depth. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For the two CelebA datasets, model depth has a negligible effect on the worst-group error whereas on the Waterbirds dataset, the increasing the model size decreases the worst-group error.

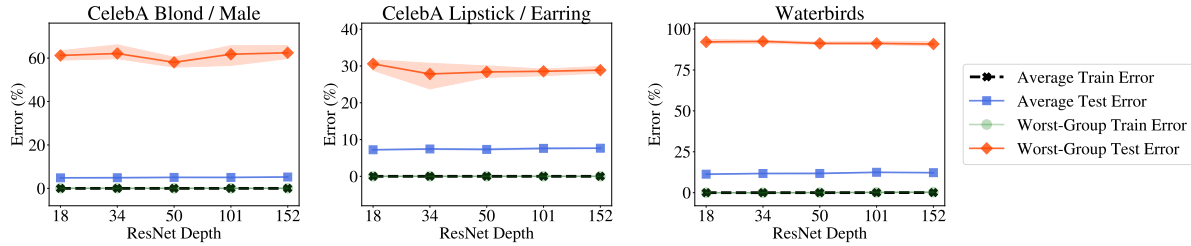


Figure 1.4: Depth of randomly initialized ResNet models is varied, increasing in depth from left to right. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For all of the datasets, model depth has a negligible effect on the worst-group error.

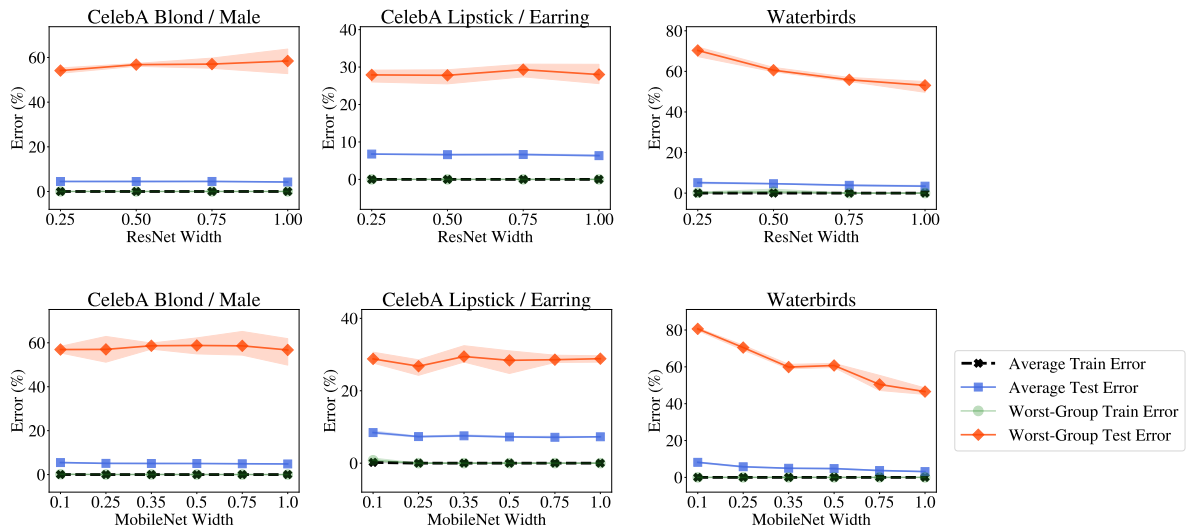


Figure 1.3: The top row shows the pre-trained ResNet18 models of varying widths. The bottom row shows the pre-trained MobileNet models of varying width. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For the two CelebA datasets, model depth has a negligible effect on the worst-group error whereas on the Waterbirds dataset, the increasing the model size decreases the worst-group error.

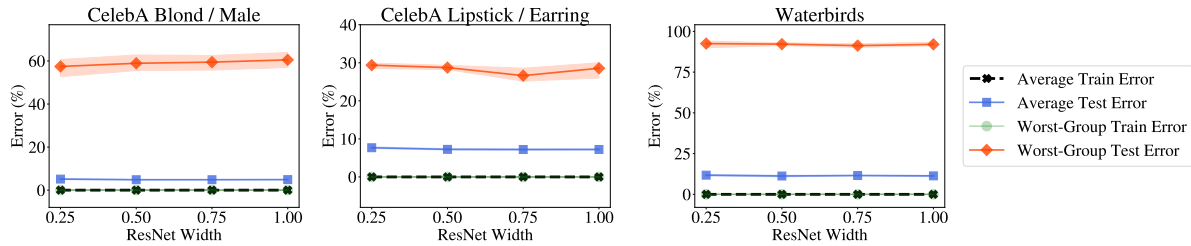


Figure 1.5: Width of randomly initialized ResNet models is varied, increasing in width from left to right. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For all of the datasets, model depth has a negligible effect on the worst-group error.

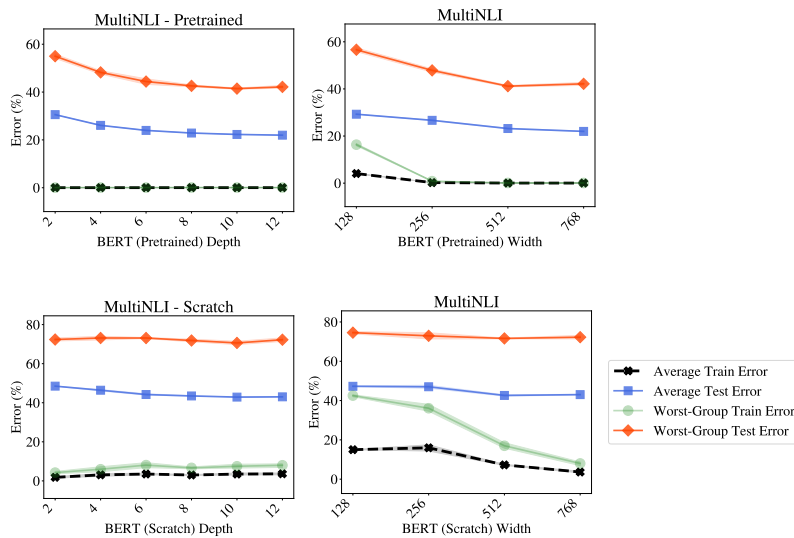


Figure 1.6: **Top Row:** Depth and width of pre-trained BERT models are varied, increasing in size from left to right. **Bottom Row:** Depth and width of randomly initialized BERT models are varied, increasing in size from left to right. Increasing pre-trained model size reduces worst-group error, while on randomly initialized models, model size has negligible effect.

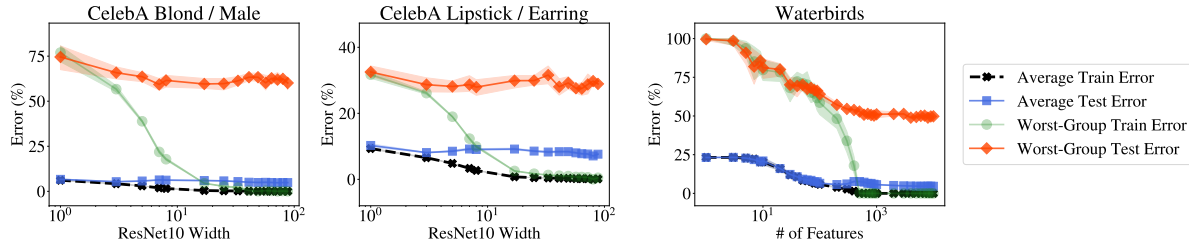


Figure 1.7: Graphs displaying how the error changes as we vary the width or number of features. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For CelebA Blond / Male and Waterbirds, worst-group error improves with the greater model size whereas worst-group error on the CelebA Lipstick / Earring seems to stay at about the same value as model size increases.

Analysis

Pre-trained Models and Randomly Initialized Models, Varying Last Layer Width.

We replicate experiments in prior work and extend results with an additional dataset to show that our setup is comparable. This set of experiments follows the setup of Sagawa et al. [25]. Performance on the Waterbirds and CelebA Blond / Male datasets gives us results consistent with prior work: worst-group error dips close to 50% for larger widths. Furthermore, we also study the CelebA Lipstick / Earring dataset and the worst-group error approaches 30% as we increase the model width. The results are summarized in Figure 1.7. Overall, we find that worst-group error decreases slightly for both CelebA datasets before saturating. On Waterbirds, the worst-group error decreases as we increase the number of features and remains around 50% when the number of features is larger than 10³.

Pre-trained Models, Varying Depth and Width. For pre-trained models, we find that model size does not hurt worst-group test error over multiple model architecture series. We compare model sizes by varying depth on ResNet, VGG BN, and BERT; and width on ResNet18, MobileNet, and BERT. On the Waterbirds and MultiNLI datasets, larger models monotonically improve worst-group accuracy. For the two CelebA datasets, increasing model sizes neither significantly increases nor decreases the worst-group test error. Results for varying depth of ResNet and VGG BN models are summarized in Figure 1.2; Resnet18 with varying width is given in Figure 1.3; and BERT results are given in Figure 1.6.

Randomly Initialized Models, Varying Depth and Width. For models trained from scratch, we also find that model size does not hurt worst-group test error. However, unlike the pre-trained results, the error did not improve with the increasing model size. ResNet results can be seen in Figure 1.4 for varying depth and Figure 1.5 for varying width, while BERT results can be seen in Figure 1.6.

Experimental Takeaways

Empirically, worst-group test error decreases or stays the same as model size increases. The experiments show that pre-trained models perform significantly better than those trained from scratch. On the Waterbirds and MultiNLI datasets we find that increasing the size of the pre-trained model improves performance on the worst group. This suggests that pretraining may be a factor in improving performance of overparameterized models.

On the other hand, we show in Appendix 1.3 that the worst-group performance of group DRO (which uses group label information) is generally better than that of ERM. Therefore, group DRO should be used when group labels are available, and our analysis primarily applies to the case where they are unavailable.

1.6 Conclusion

Although increasing model size only sometimes helps worst-group generalization, large models generally do not hurt across almost all the ERM settings, whether the model is pre-trained or trained from scratch. Furthermore, we find that as compared to models trained from scratch, increasing pre-trained model size is more likely to improve worst-group accuracy. We leave for future work the effects of pre-training on the worst-group performance. Our experimental results suggest that the study of spurious correlations under the ERM setting is interesting from both the practical and analytical perspectives, and it can potentially lead to novel ways of designing experimental protocols and algorithms.

Bibliography

- [1] Mikhail Belkin et al. *Reconciling modern machine learning practice and the bias-variance trade-off*. 2019. arXiv: 1812.11118 [stat.ML].
- [2] Jorg Bornschein, Francesco Visin, and Simon Osindero. *Small Data, Big Decisions: Model Selection in the Small-Data Regime*. 2020. arXiv: 2009.12583 [cs.LG].
- [3] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 77–91. URL: <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- [4] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: <http://arxiv.org/abs/1810.04805>.
- [5] Karan Goel et al. *Model Patching: Closing the Subgroup Performance Gap with Data Augmentation*. 2020. arXiv: 2008.06775 [cs.LG].
- [6] Suchin Gururangan et al. “Annotation Artifacts in Natural Language Inference Data”. In: *CoRR* abs/1803.02324 (2018). arXiv: 1803.02324. URL: <http://arxiv.org/abs/1803.02324>.
- [7] Tatsunori Hashimoto et al. “Fairness Without Demographics in Repeated Loss Minimization”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1929–1938. URL: <https://proceedings.mlr.press/v80/hashimoto18a.html>.
- [8] Tatsunori B. Hashimoto et al. *Fairness Without Demographics in Repeated Loss Minimization*. 2018. arXiv: 1806.08010 [stat.ML].
- [9] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: <http://arxiv.org/abs/1512.03385>.
- [10] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: <http://arxiv.org/abs/1512.03385>.
- [11] Dan Hendrycks et al. *Natural Adversarial Examples*. 2021. arXiv: 1907.07174 [cs.LG].

- [12] Katherine L. Hermann, Ting Chen, and Simon Kornblith. *The Origins and Prevalence of Texture Bias in Convolutional Neural Networks*. 2020. arXiv: 1911.09071 [cs.CV].
- [13] Divyansh Kaushik, Eduard H. Hovy, and Zachary C. Lipton. “Learning the Difference that Makes a Difference with Counterfactually-Augmented Data”. In: *CoRR* abs/1909.12434 (2019). arXiv: 1909.12434. URL: <http://arxiv.org/abs/1909.12434>.
- [14] Fereshte Khani and Percy Liang. *Removing Spurious Features can Hurt Accuracy and Affect Groups Disproportionately*. 2020. arXiv: 2012.04104 [cs.LG].
- [15] Pang Wei Koh et al. “WILDS: A Benchmark of in-the-Wild Distribution Shifts”. In: *CoRR* abs/2012.07421 (2020). arXiv: 2012.07421. URL: <https://arxiv.org/abs/2012.07421>.
- [16] Duo Li, Aojun Zhou, and Anbang Yao. *HBONet: Harmonious Bottleneck on Two Orthogonal Dimensions*. 2019. arXiv: 1908.03888 [cs.CV].
- [17] Evan Zheran Liu et al. *Just Train Twice: Improving Group Robustness without Training Group Information*. 2021. arXiv: 2107.09044 [cs.LG].
- [18] Ziwei Liu et al. “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*. 2015.
- [19] Song Mei and Andrea Montanari. *The generalization error of random features regression: Precise asymptotics and double descent curve*. 2020. arXiv: 1908.05355 [math.ST].
- [20] Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. “Overparameterisation and worst-case generalisation: friend or foe?” In: *ICLR*. 2021.
- [21] Preetum Nakkiran et al. “Deep Double Descent: Where Bigger Models and More Data Hurt”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: <https://openreview.net/forum?id=B1g5sA4twr>.
- [22] Timothy Niven and Hung-Yu Kao. “Probing Neural Network Comprehension of Natural Language Arguments”. In: *CoRR* abs/1907.07355 (2019). arXiv: 1907.07355. URL: <http://arxiv.org/abs/1907.07355>.
- [23] Luke Oakden-Rayner et al. “Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging”. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. CHIL ’20. Toronto, Ontario, Canada: Association for Computing Machinery, 2020, 151–159. ISBN: 9781450370462. DOI: 10.1145/3368555.3384468. URL: <https://doi.org/10.1145/3368555.3384468>.
- [24] Alan Pham et al. “The Effect of Model Size on Worst-Group Generalization”. In: *NeurIPS DistShift Workshop*. 2021.
- [25] Shiori Sagawa et al. “An Investigation of Why Overparameterization Exacerbates Spurious Correlations”. In: *ICML*. 2020.

- [26] Shiori Sagawa et al. *Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization*. 2020. arXiv: 1911.08731 [cs.LG].
- [27] Mark Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: 1801.04381 [cs.CV].
- [28] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [29] Lifu Tu et al. “An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models”. In: *TACL* 8 (2020).
- [30] Iulia Turc et al. “Well-Read Students Learn Better: On the Importance of Pre-training Compact Models”. In: *arXiv preprint arXiv:1908.08962v2* (2019).
- [31] V Vapnik. “Statistical learning theory new york”. In: *NY: Wiley* 1 (1998), p. 2.
- [32] Tan Wang et al. *Causal Attention for Unbiased Visual Recognition*. 2021. arXiv: 2108.08782 [cs.CV].
- [33] P. Welinder et al. *Caltech-UCSD Birds 200*. Tech. rep. CNS-TR-2010-001. California Institute of Technology, 2010.
- [34] Adina Williams, Nikita Nangia, and Samuel Bowman. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1112–1122. URL: <http://aclweb.org/anthology/N18-1101>.
- [35] Zitong Yang et al. *Rethinking Bias-Variance Trade-off for Generalization of Neural Networks*. 2020. arXiv: 2002.11328 [cs.LG].
- [36] Ruiqi Zhong et al. *Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level*. 2021. arXiv: 2105.06020 [cs.CL].
- [37] Bolei Zhou et al. “Places: A 10 million Image Database for Scene Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).
- [38] Chunting Zhou et al. *Examining and Combating Spurious Features under Distribution Shift*. 2021. arXiv: 2106.07171 [cs.LG].