The Effect of Model Size on Worst-Group Generalization

Alan Pham*, **Eunice Chan***, **Vikranth Srivatsa***, **Dhruba Ghosh***,
Yaoqing Yang, **Yaodong Yu**, **Ruiqi Zhong**, **Joseph E. Gonzalez**†, **Jacob Steinhardt**†
Department of Electrical Engineering and Computer Science
University of California, Berkeley
{alanlp, ekchan, vsrivatsa, djghosh13, yqyang, yaodong_yu,
ruiqizhong, jegonzal, jsteinhardt}@berkeley.edu

Abstract

Overparameterization is often believed to hurt test accuracy on rare subgroups. However, prior works establishing this focused on cases where subgroup information is known and group DRO is applied to improve the performance of the worst subgroup. 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 error under ERM across all setups. In particular, larger pre-trained models are consistently better on Waterbirds and MultiNLI. We advise practitioners to use larger pre-trained models when subgroup labels are unknown.

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 (Bornschein et al., 2020). Despite this trend, many works find or assume that overparameterization can hurt test accuracy on certain subgroups of the data (Buolamwini & Gebru, 2018; Hashimoto et al., 2018a; Menon et al., 2021; Sagawa et al., 2020a; Zhong et al., 2021). This problem is exacerbated when tuning a model with unknown group labels since the performance gap between average accuracy and worst-group accuracy can be large. The difference between the average test accuracy and worst-group test accuracy ranges from 60 to 40% in the empirical risk minimization (ERM) setup from Sagawa et al. (2020b) 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.

A model might suffer more from distribution shifts if it relies on spurious features — features that correlate with the class label in the training dataset, but have no casual relationship. 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. In this paper, we are interested in model performance on the rare subgroups where the spurious features hinders rather than helps prediction, e.g., in the

Both * and † mean equal contribution.

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

setting introduced above would be the subgroups of waterbirds on land and landbirds on water. These spurious correlations are an obstacle to solving real world problems: thoracic pathology detection models are found to rely on the presence of chest drains (a treatment device) to detect pneumothorax (Oakden-Rayner et al., 2020).

We systematically investigate the effect of model size on the performance of the model on the rare subgroups. Specifically, we work with rare subgroups that cannot use the spurious cues for prediction. Prior work generally show that increasing model size can hurt worst-group performance. However, these works focus on reweighing methods such as group DRO. A supplementary result in Sagawa et al. (2020b) shows that ERM error improves or at least does not get worse as model size increases. In our work, we provide complementary findings as we explore this trend 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) (Sagawa et al., 2020a), CelebA (a face dataset we use for two tasks: Lipstick / Earring and Blond / Male) (Liu et al., 2015), and MultiNLI (a language dataset) (Williams et al., 2018) datasets (more details in Appendix B). 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 4). 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 2).

Interestingly, when comparing the performance of models that were pretrained and models that were trained from scratch, we notice worst-group error shows a downward

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 number 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.

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 is also found to improve robustness to adversarial samples and distributional shifts (Hendrycks et al., 2021). This has been attributed to the "double descent" phenomenon, which explains the increase in accuracy from significantly overparameterizing models (Belkin et al., 2019). Bornschein et al. (2020) discusses this in the context of small data, while Yang et al. (2020) provides an explanation via the bias-variance trade-off curve. Sagawa et al. (2020b) 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. 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) (Tu et al., 2020). On the other hand, Niven & Kao (2019) 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 (Zhou et al., 2021; Hermann et al., 2020).

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 (Kaushik et al., 2019). 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. (2020) suggests using CycleGAN to generate negative examples. Liu et al. (2021) suggests a two-stage training regime that upweights misclassified points. Wang et al. (2021) 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 (Zhou et al., 2021; Hashimoto et al., 2018b). One approach is group DRO (GDRO), an instance of DRO that minimizes the worst-group expected loss (Sagawa et al., 2020a, Eqn.(4)). This substantially improves generalization when coupled with heavy regularization, but requires prior labeling of subgroups in the training set.

3 Problem Setting

We adopt the formulation presented in Sagawa et al. (2020b). 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) (Vapnik, 1998), finding the model parameters θ that empirically minimize the average training loss:

$$\mathcal{R}_{\text{ERM}}(\theta) = \mathbb{E}_{(x,y,g)}[\ell(\theta, (x, y))]. \tag{1}$$

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 B.3.

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

$$\varepsilon_g := \mathbb{E}_{x,y|g}[\mathbf{1}(h(x) \neq y)]. \tag{2}$$

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:

$$\varepsilon := \sum_{g \in G} w_g \times \varepsilon_g \quad \text{and} \quad \varepsilon_{\text{wg}} := \max_{g \in \mathcal{G}} \varepsilon_g \tag{3}$$

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.

4 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.

4.1 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

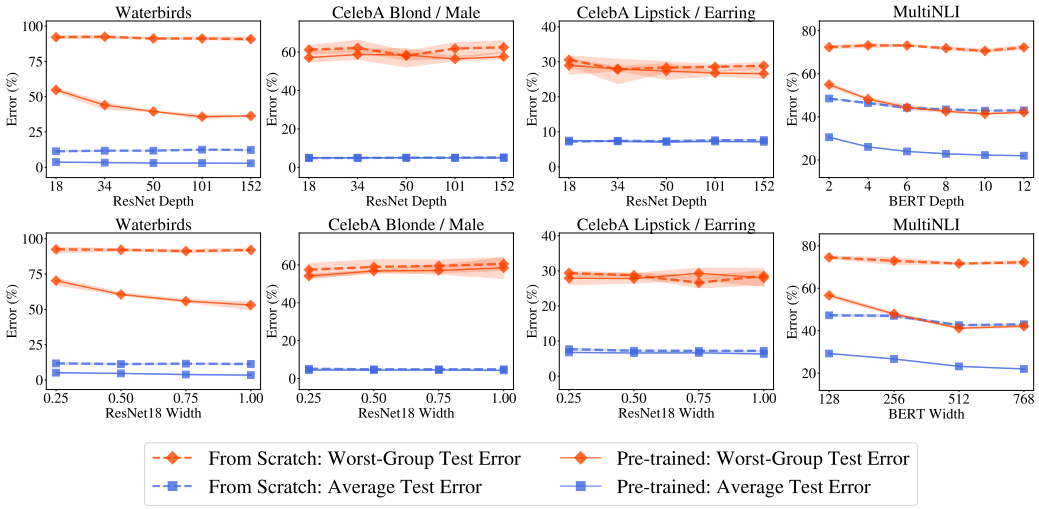

Figure 1: Models trained to convergence, e.g., until average training accuracy is close or equal to 100%. Hyperparameters remain the same within each experiment series. In each of the four graphs above, we compare the error of pre-trained and randomly initialized models of the same architecture. Pre-trained 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.

comparable. This set of experiments follows the setup of Sagawa et al. (2020b). 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 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 10^3.

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 2; Resnet18 with varying width is given in Figure 3; and BERT results are given in Figure 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 4 for varying depth and Figure 5 for varying width, while BERT results can be seen in Figure 6.

4.2 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 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.

5 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.

Acknowledgments

We would like to thank Pang Wei Koh and Aditi Raghunathan for their valuable feedback and comments during preparation of this manuscript. Joseph E. Gonzalez would like to acknowledge supports from NSF CISE Expeditions Award CCF-1730628 and gifts from Alibaba, Amazon Web Services, Ant Group, Ericsson, Facebook, Futurewei, Google, Intel, Microsoft, Nvidia, Scotiabank, Splunk and VMware. Our conclusions do not necessarily reflect the position or the policy of our sponsors, and no official endorsement should be inferred.

References

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off, 2019. 2

Jorg Bornschein, Francesco Visin, and Simon Osindero. Small data, big decisions: Model selection in the small-data regime, 2020. 1, 2

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html. 1

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805. 10

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation, 2020. 3

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. *CoRR*, abs/1803.02324, 2018. URL http://arxiv.org/abs/1803.02324. 11

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/hashimoto18a.html. 1

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization, 2018b. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015a. URL http://arxiv.org/abs/1512.03385. 10

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015b. URL http://arxiv.org/abs/1512.03385. 10

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples, 2021. 2

Katherine L. Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks, 2020. 2

Divyansh Kaushik, Eduard H. Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *CoRR*, abs/1909.12434, 2019. URL http://arxiv.org/abs/1909.12434. 3

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately, 2020. 11

Duo Li, Aojun Zhou, and Anbang Yao. Hbonet: Harmonious bottleneck on two orthogonal dimensions, 2019. 11

Evan Zheran 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, 2021. 3

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 11

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve, 2020. 10

Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *ICLR*, 2021. 1

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 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. 10

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. *CoRR*, abs/1907.07355, 2019. URL http://arxiv.org/abs/1907.07355. 2

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. 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, pp. 151–159, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384468. URL https://doi.org/10.1145/3368555.3384468. 2

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, 2020a. 1, 2, 3, 11, 12

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, 2020b. 1, 2, 3, 4

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. 10, 11

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 10

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *TACL*, 8, 2020. 2

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*, 2019. 11

V Vapnik. Statistical learning theory new york. *NY: Wiley*, 1:2, 1998. 3

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition, 2021. 3

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 11

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)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101. 2, 11

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks, 2020. 2

Ruiqi Zhong, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. Are larger pretrained language models uniformly better? comparing performance at the instance level, 2021. 1

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 11

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift, 2021. 2, 3

A Appendix

We summarize our depth varying for pre-trained models results in figure 2, width varying for pre-trained models in figure 3, depth varying for randomly initialized models in figure 4, and width varying for randomly initialized models in figure 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.

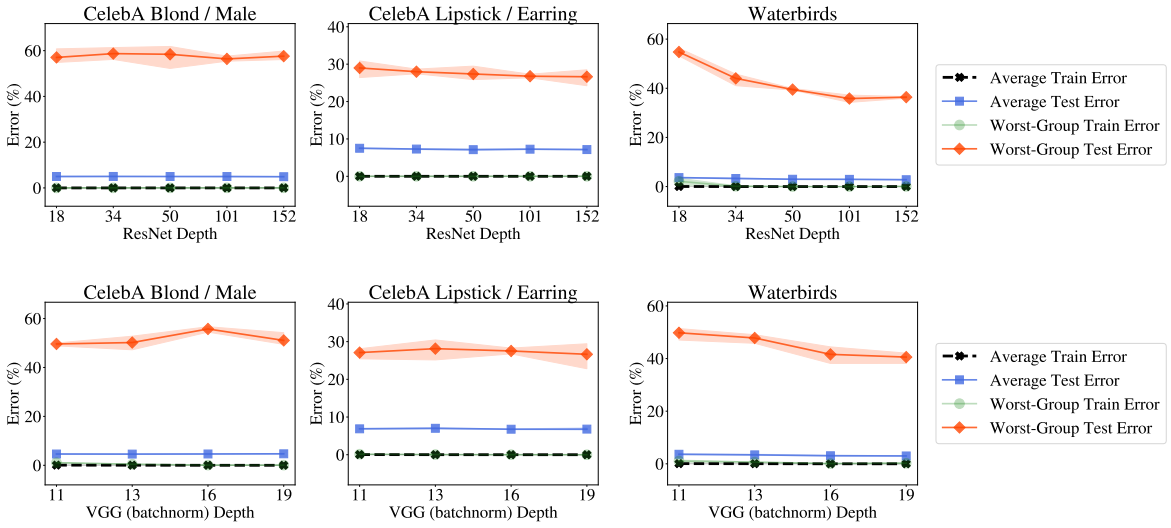

Figure 2: The top row shows the pre-trained ResNet models of varying depth. The bottom row shows the pre-trained VGG models of varying depth. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For 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.

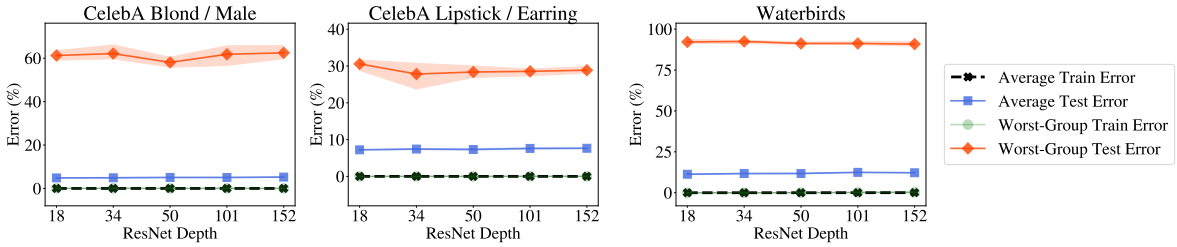

Figure 4: Depth of randomly initialized ResNet models is varied, increasing in depth from left to right. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For all of the datasets, model depth has a negligible effect on the worst-group error.

Figure 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.

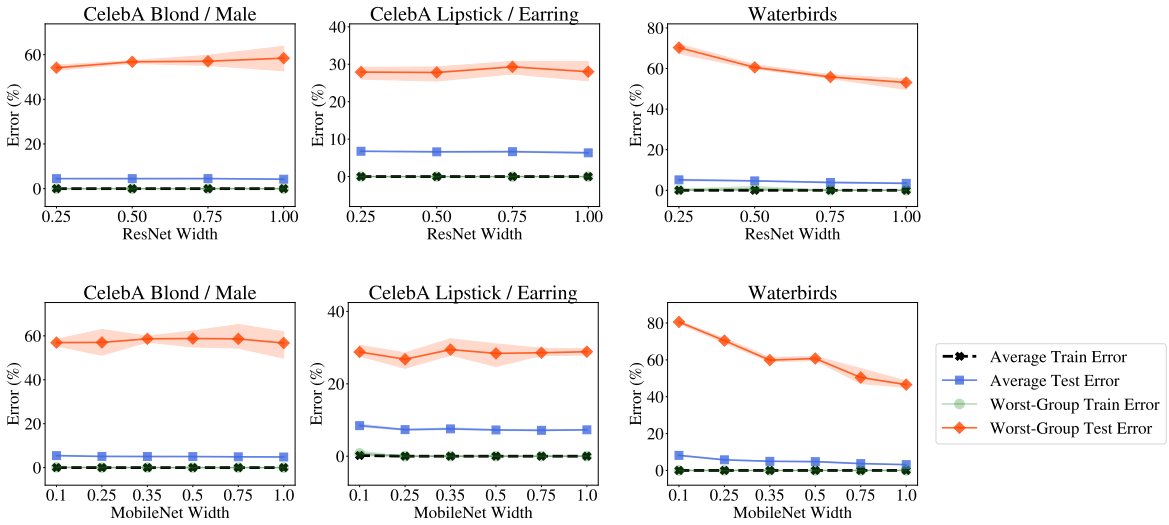

Figure 3: The top row shows the pre-trained ResNet18 models of varying widths. The bottom row shows the pre-trained MobileNet models of varying width. Each column represents the dataset the model is trained and evaluated on. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. 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.

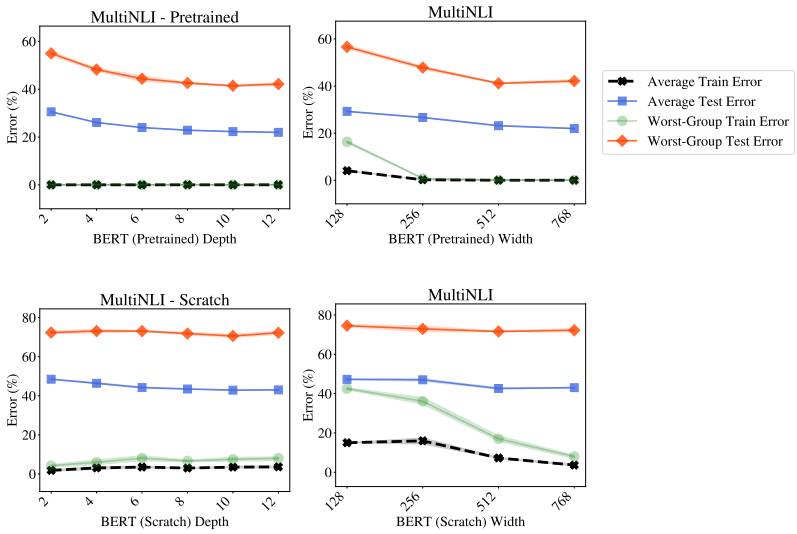

Figure 6: **Top Row:** Depth and width of pre-trained BERT models are varied, increasing in size from left to right. **Bottom Row:** Depth and width of randomly initialized BERT models are varied, increasing in size from left to right. Increasing pre-trained model size reduces worst-group error, while on randomly initialized models, model size has negligible effect.

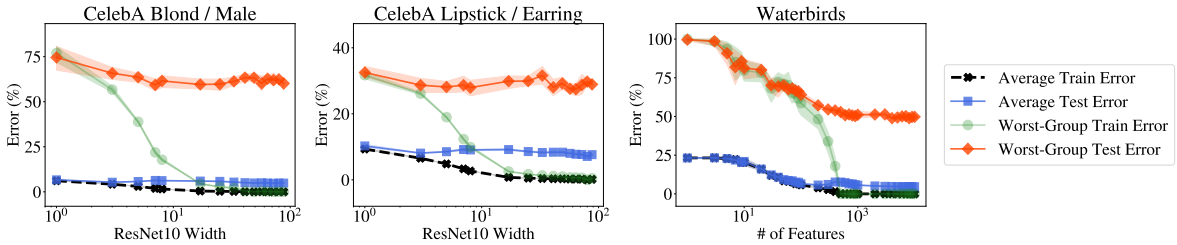

Figure 7: Graphs displaying how the error changes as we vary the width or number of features. From left to right: CelebA Blond / Male, CelebA Lipstick / Earring, and Waterbirds. For CelebA Blond / Male and Waterbirds, worst-group error improves with the greater model size whereas worst-group error on the CelebA Lipstick / Earring seems to stay at about the same value as model size increases.

B 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 (He et al., 2015b), VGG (batchnorm) (Simonyan & Zisserman, 2014), and MobileNet (Sandler et al., 2019). For the NLP domain, we use BERT (Devlin et al., 2018). 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. (2020), varying the width of a ResNet10 model (He et al., 2015a). The setup over the Waterbirds dataset follows the setup in Mei & Montanari (2020), 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.

B.1 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 (Sagawa et al., 2020a) by cropping out bird photographs from the Caltech-UCSD Birds-200-2011 (CUB) dataset (Welinder et al., 2010) and placing them on top of image backgrounds from the Places dataset (Zhou et al., 2017). 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 (Liu et al., 2015) is a large scale multi-feature face dataset with varying backgrounds and poses. Using the CelebA dataset, we can construct spurious correlations 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 (Khani & Liang, 2020). 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 Williams et al. (2018). 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, Gururangan et al. (2018) 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 B.3.1 in the appendix describes the number of samples in each group.

B.2 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 (Sandler et al., 2019). We use architectures and pretrained checkpoints from a Github reproduction from Li et al. (2019).

BERT. We use the BERT architecture for the MultiNLI task, varying the width and depth as in Turc et al. (2019), which showed effectiveness of BERT even with non-standard depths and widths.

B.3 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 2.

B.3.1 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 Count	Worst-Group Count
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: Number of training examples in each dataset

Dataset	Inital 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 2: The hyperparameters used to train the models to convergence on each dataset.

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

Table 3: Comparison of group DRO summarized from Sagawa et al. (2020a) with our own ERM results.