Understanding Rare Spurious Correlations in Neural Networks

Yao-Yuan Yang [1] **Chi-Ning Chou** [2] **Kamalika Chaudhuri** [1]

Abstract

Neural networks are known to use spurious correlations such as background information for classification. While prior work has looked at spurious correlations that are widespread in the training data, in this work, we investigate how sensitive neural networks are to *rare* spurious correlations, which may be harder to detect and correct, and may lead to privacy leaks. We introduce spurious patterns correlated with a fixed class to a few training examples and find that it takes only a handful of such examples for the network to learn the correlation. Furthermore, these rare spurious correlations also impact accuracy and privacy. We empirically and theoretically analyze different factors involved in rare spurious correlations and propose mitigation methods accordingly. Specifically, we observe that ℓ_2 regularization and adding Gaussian noise to inputs can reduce the undesirable effects.

1. Introduction

Neural networks are known to use spurious patterns for classification. Image classifiers use background as a feature to classify objects (Gururangan et al., 2018; Srivastava et al., 2020; Zhou et al., 2021) often to the detriment of generalization (Nagarajan et al., 2020). For example, Sagawa et al. (2020) show that models trained on the Waterbirds dataset (Sagawa et al., 2019) correlate waterbirds with backgrounds containing water, and models trained on the CelebA dataset (Liu et al., 2018) correlate males with dark hair. In all these cases, spurious patterns are present in a substantial number of training points. The vast majority of waterbirds, for example, are photographed next to the water.

Understanding how and when spurious correlations appear in neural networks is a frontier research problem and remains elusive. In this paper, we study spurious correlations in the context where the appearance of spurious patterns is *rare* in the training data. Our motivations are three-fold. First, while it is reasonable to expect that widespread spurious correlations in the training data will be learnt, a related question is what happens when these correlations are *rare*. Understanding if and when they are learnt and how to mitigate them is a first and necessary step before we can understand and mitigate spurious correlations more broadly. Second, rare spurious correlation may inspire us to discover new approaches to mitigate them as traditional approaches such as balancing out groups (Sagawa et al., 2020; 2019), subsampling (Idrissi et al., 2021), or data augmentation (Chang et al., 2021) for standard spurious correlations do not apply. Third, rare spurious correlations naturally connect to data privacy. For example, in Leino & Fredrikson (2020), the training set had an image of Tony Blair with a pink background. This led to a classifier that assigned a higher likelihood of the label "Tony Blair" to images with a pink background. Thus, an adversary could exploit this to infer the existence of "Tony Blair" with a pink background in the training set by by presenting images of other labels with a pink background.

We systematically investigate rare spurious correlations through the following three research questions. First, when do spurious correlations appear, i.e., how many training points with the spurious pattern would cause noticeable spurious correlations? Next, how do rare spurious correlations affect neural networks? Finally, is there any way to mitigate the undesirable effects of rare spurious correlations?

Overview

We attempt to answer the above questions via both experimental and theoretical approaches. On the experimental side, we introduce spurious correlations into real image datasets by turning a few training data into *spurious examples*, i.e., adding a spurious pattern to a training image from a target class. We then train a neural network on the modified dataset and measure the strength of the correlation between the spurious pattern and the target class in the network. On the theoretical side, we design a toy mathematical model that enables quantitative analysis on different factors (e.g., the fraction of spurious examples, signal-to-noise ra-

[1]University of California San Diego [2]Harvard University. Correspondence to: Yao-Yuan Yang <yay005@eng.ucsd.edu>, Chi-Ning Chou <chiningchou@g.harvard.edu>, Kamalika Chaudhuri <kamalika@eng.ucsd.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

tio, etc.) of rare spurious correlations. Our responses to the three research questions are summarized in the following.

Rare spurious correlations appear even when the number of spurious samples is small. Empirically, we define a *spurious score* to measure the amount of spurious correlations. We find that the spurious score of a network trained with only 1 spurious examples out of 60,000 training samples can be significantly higher than that of the baseline. A visualization of the model also reveals that the network's weights may be significantly affected by the spurious pattern. In our theoretical model, we further discover that there is a sharp phase transition of spurious correlations from no spurious training example to a non-zero fraction of spurious training examples. Together, these findings provide a strong evidence that spurious correlations can be learnt even when the number of spurious samples is extremely small.

Rare spurious correlations affect both the privacy and test accuracy. We analyze the privacy issue of rare spurious correlations via the membership inference attack (Shokri et al., 2017; Yeom et al., 2017), which measures the privacy level according to the hardness of distinguishing training samples from testing samples. We observe that the spurious training examples are more vulnerable to membership inference attacks. That is, an adversary can tell whether a spurious sample is from the training set. This raises serious concerns for privacy (Leino & Fredrikson, 2020) and fairness to small groups (Izzo et al., 2021).

We examine the effect of rare spurious correlations on test accuracy through two accuracy notions: the clean test accuracy, which uses the original test examples, and the spurious test accuracy, which adds the spurious pattern to all the test examples. Both empirically and theoretically, we find that clean test accuracy does not change too much while the spurious test accuracy significantly drops in the face of rare spurious correlations. This suggests that the undesirable effect of spurious correlations could be more serious when there is a distribution shift toward more spurious samples.

Methods to mitigate the undesirable effects of rare spurious correlations. Finally, inspired by our theoretical analysis, we examine three regularization methods to reduce the privacy and test accuracy concerns: adding Gaussian noises to the input samples, ℓ_2 regularization (or equivalently, weight decay), and gradient clipping. We find that adding Gaussian noise and ℓ_2 regularization effectively reduce spurious score and improve spurious test accuracy. Meanwhile, not all regularization methods could reduce the effects of rare spurious correlations, e.g., gradient clipping. Our findings suggest that rare spurious correlations should be dealt differently from traditional privacy issues. We post it as a future research problem to deepen the understanding of how to mitigate rare spurious correlations.

Concluding remarks. The study of spurious correlations is crucial for a better understanding of neural networks. In this work, we take a step forward by looking into a special (but necessary) case of spurious correlations where the appearance of spurious examples is rare. We demonstrate both experimentally and theoretically when and how rare spurious correlations appear and what undesirable consequences are. While we propose a few methods to mitigate rare spurious correlations, we emphasize that there is still a lot to explore, and we believe the study of rare spurious correlations could serve as a guide for understanding the more general cases.

2. Rare Spurious Correlations are Learnt

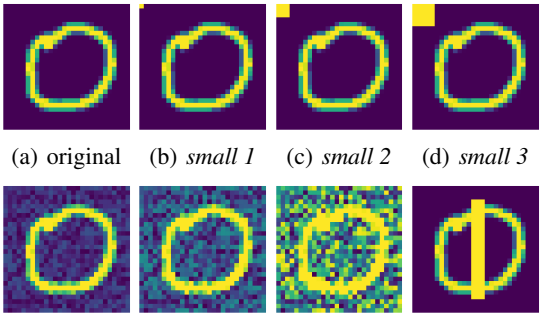
We start with an empirical study of rare spurious correlations in neural networks. We train a neural network using a modified training dataset given by the *overlapping model* where a spurious pattern is added to a few training examples with the same label (target class). We then analyze the effect of these spurious training examples through three difference angles: (i) a quantitative analysis on the appearance of spurious correlations via an empirical measure, *spurious score*, (ii) a qualitative analysis on the appearance of spurious correlations through visualizing the network weights, and (iii) an analysis on the consequences of rare spurious correlations in terms of privacy and test accuracy.

2.1. Introducing spurious examples to networks

As we don't have access to the underlying ground-truth feature of an empirical data, we artificially introduce spurious features into the training dataset. Concretely, given a dataset (e.g., MNIST), we treat each training example $\mathbf{x}$ as an invariant feature. Next, we pick a target class c_{tar} (e.g., the zero class), a spurious pattern $\mathbf{x}_{\text{sp}}$ (e.g., a yellow square at the top-left corner), and a mapping $\Phi_{\mathcal{X}}$ that combines a training example with the spurious pattern. Finally, we randomly select n training examples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from the target class c_{tar} and replace these examples with $\Phi_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_{\text{sp}})$ for each $i = 1, \ldots, n$. See Fig. 1 and the following paragraphs and Appendix E.1 for detailed experiment setups.

Datasets & the target class c_{tar}. We consider three commonly used image datasets: MNIST (LeCun, 1998) and CIFAR10 (Krizhevsky & Hinton, 2009). MNIST have $60,000$ training examples, and CIFAR10 has $50,000$. We set the first two classes of each dataset as the target class ($c_{\text{tar}} = \{0, 1\}$), which are zero and one for MNIST and airplane and automobile for CIFAR10.

Spurious patterns $\mathbf{x}_{\text{sp}}$. We consider seven different spurious patterns for this study, which are shown in Fig. 1. The patterns *small 1* (*S1*), *small 2* (*S2*), and *small 3* (*S3*) are designed to test if a neural network can learn the correlations



(a) original (b) *small 1* (c) *small 2* (d) *small 3*

(e) *random 1* (f) *random 2* (g) *random 3* (h) *invariant*

Figure 1. Different spurious patterns considered in the experiment.

between small patterns and the target class. The patterns *random 1* (*R1*), *random 2* (*R2*), and *random 3* (*R3*) are patterns with each pixel value being uniformly random sampled from $[0, r]$, where $r = 0.25, 0.5, 1.0$. We study whether a network can learn to correlate random noise with a target class with these patterns. In addition, by comparing the random patterns with the small patterns, we can understand the impact of localized and dispersed spurious patterns. Lastly, the pattern *core* (*Inv*) is designed for MNIST with $c_{\text{tar}} = 0$ to understand what happens if the spurious pattern overlaps with the core feature of another class.

Combination function $\Phi_{\mathcal{X}}$. The function $\Phi_{\mathcal{X}}$ combines the original example $\mathbf{x}$ with the spurious pattern $\mathbf{x}_{\text{sp}}$ into a spurious example. For simplicity, we consider the *overlapping model* where $\Phi_{\mathcal{X}}$ directly adds the spurious pattern $\mathbf{x}_{\text{sp}}$ onto the original example $\mathbf{x}$ and then clips the value of each pixel to $[0, 1]$, i.e., $\Phi_{\mathcal{X}}(\mathbf{x}, \mathbf{x}_{\text{sp}}) = clip_{[0,1]}(\mathbf{x} + \mathbf{x}_{\text{sp}})$.

The number of spurious examples. For MNIST, we randomly insert the spurious pattern to $n = 0, 1, 3, 5, 10, 20, 100, 2000$, and 5000 training examples labeled as the target class c_{tar}. These training examples inserted with a spurious pattern are called spurious examples. For CIFAR10, we consider datasets with $n = 0, 1, 3, 5, 10, 20, 100,\ 500,$ and 1000 spurious examples. Note that 0 spurious example means the original training set is not modified.

2.2. Quantitative analysis: spurious score

To evaluate the strength of spurious correlations in a neural network, we design an empirical quantitative measure, *spurious score*, as follows. Let $f_c(\mathbf{x})$ be the neural network's predicted probability of an example $\mathbf{x}$ belonging to class c. Intuitively, the larger the *prediction difference* $f_{c_{\text{tar}}}(\mathbf{x}) - f_{c_{\text{tar}}}(\Phi_{\mathcal{X}}(\mathbf{x}, \mathbf{x}_{\text{sp}}))$ is, the stronger spurious correlations the neural network f had learned. To quantify the effect of spurious correlations, we measure how frequently the prediction difference of the test examples exceed a certain threshold. Formally, let $\epsilon > 0$, we define the *ϵ-spurious score* as the fraction of test example $\mathbf{x}$ that satisfies

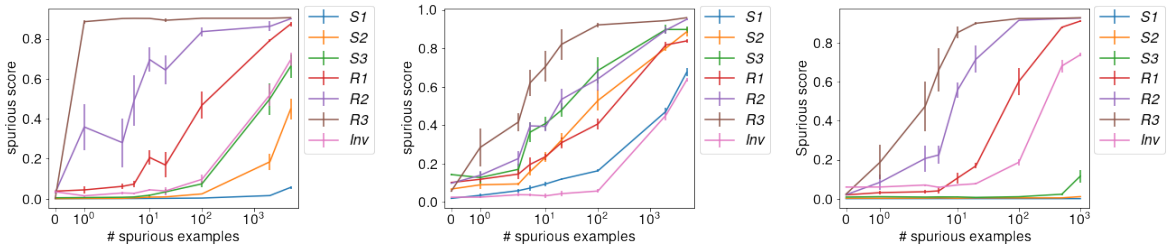
$$f_{c_{\text{tar}}}(\Phi_{\mathcal{X}}(\mathbf{x}, \mathbf{x}_{\text{sp}})) - f_{c_{\text{tar}}}(\mathbf{x}) > \epsilon\,. \tag{1}$$

In other words, spurious score measures the portion of test examples that get an non-trivial increase in the predicted probability of the target class c_{tar} when the spurious pattern is presented. We make three remarks on the definition of spurious score. First, as we don't have any prior knowledge on the structure of f, we use the fraction of test examples satisfying Eq. (1) as opposed to other function of $f_{c_{\text{tar}}}(\Phi_{\mathcal{X}}(\mathbf{x}, \mathbf{x}_{\text{sp}})) - f_{c_{\text{tar}}}(\mathbf{x})$ (e.g., taking average) to avoid non-monotone or unexplainable scaling. Second, the choice of the threshold ϵ is to avoid numerical errors to affect the result. In our experiment, we pick $\epsilon = 1/(\#\text{classes})$ (e.g., in MNIST we pick $\epsilon = 1/10$) and empirically similar conclusions can be made with other choices of ϵ. Finally, we point out that spurious score captures the privacy concern raised by the "Tony Blair" example mentioned in the introduction.

Empirical findings. We repeat the measurement of spurious scores on five neural networks trained with different random seeds and summarize the results in Fig. 2. Fig. 2 shows the spurious scores for each dataset and pattern as a function of the number of spurious examples. Starting with the random pattern *R3*, we see that the spurious scores increase significantly from zero to three spurious examples in all six cases (three datasets and two target classes). This shows that neural networks can learn rare spurious correlations with as little as *one to three spurious examples*. Since all three datasets have $50,000$ or more training examples, it is surprising that the networks learn a strong correlation with extremely small amount of spurious examples.

A closer look at Fig. 2 reveals a few other interesting observations. First, comparing the small and random patterns, we see that random patterns generally have a higher spurious score. This suggests that dispersed patterns that are spread out over multiple pixels may be more easily learnt than more concentrated ones. Second, spurious correlations are learnt even for *Inv*, on $c_{\text{tar}} = 0$ and MNIST (recall that *Inv* is designed to be similar to the core feature of class one.) This suggests that spurious correlations may be learnt even when the pattern overlaps with the foreground. Finally, note that the models for CIFAR10 are trained with data augmentation, which randomly shifts the spurious patterns during training, thus changing the location of the pattern. This suggests that these patterns can be learnt regardless of data augmentation.

Moreover, in Appendix A, we show that the neural networks weights are significantly altered even when there is a small number of spurious examples in the training set.



(a) MNIST, $c_{\text{tar}} = 0$ (b) Fashion, $c_{\text{tar}} = 0$ (c) CIFAR10, $c_{\text{tar}} = 0$

Figure 2. Each figure shows the mean and standard error of the spurious scores on three datasets, MNIST, Fashion, and CIFAR10, $c_{\text{tar}} = 0$, and different numbers of spurious examples.

2.3. Consequences of rare spurious correlations

In the previous analysis, we demonstrated that spurious correlations appear quantitatively and qualitatively in neural networks even when the number of spurious examples is small. Now, we investigate the potentially undesirable effects through the lens of privacy and test accuracy.

Privacy. We evaluate the privacy of a neural network (the target model) through membership inference attack. We follow the setup for black-box membership inference attack (Shokri et al., 2017; Yeom et al., 2017). We record how well an attack model can distinguish whether an example is from the training or testing set using the output of the target model (equivalently to a binary classification problem). If the attack model has a high accuracy, this means that the target model is leaking out information from the training (private) data. The experiment is repeated ten times. More details are in Appendix E.2.

Results on membership inference attack. Fig. 3 shows the mean and standard error of the attack model's test accuracy on all test and spurious examples. We see that the accuracies on spurious examples is generally higher when the number of spurious examples are small, which means that spurious examples are more vulnerable to membership inference attacks when appeared rarely. Although membership inference attack is a different measure for privacy than spurious score, it can be a corroboration evidence that supports the fact that privacy is leaked from spurious examples.

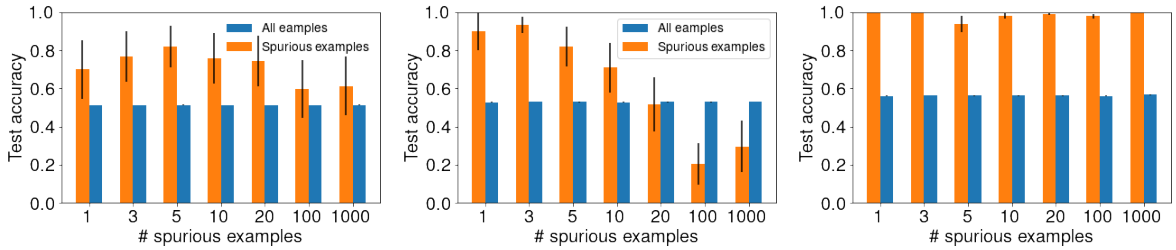
Test accuracy. We measure two types of test accuracy on neural networks trained on different number of spurious examples. The *clean test accuracy* measures the accuracy of the trained model on the original test data. The *spurious test accuracy* simulates the case where there is a distribution shift during the test time. Formally, spurious test accuracy is defined as the accuracy on a new test dataset constructed by adding spurious features to all the test examples with a label different from c_{tar}.

Results on clean test accuracy. We observe that the change in clean test accuracy in our experiments is small. Across all the models trained in Fig. 2, the minimum, maximum, average, and standard deviation of the test accuracy for each dataset are: MNIST: $(.976, .983, .980, .001)$, Fashion: $(.859, .903, .890, .010)$, CIFAR10: $(.876, .893, .886, .003)$.

Results on spurious test accuracy. The results are shown in Fig. 4. We have two observations. First, we see that there are already some accuracy drop even when spurious test accuracy is evaluated on models trained on zero spurious examples. This means that these models are not robust to the existence of spurious features. This phenomena is prominent for spurious patterns with larger norm such as *R3*. Second, we see that spurious test accuracies start to drop even more at around 10 to 100 spurious examples. This indicates that even with .01 % to .001 % of the overall training data filled with spurious examples of a certain class, the robustness to spurious features can drop significantly.

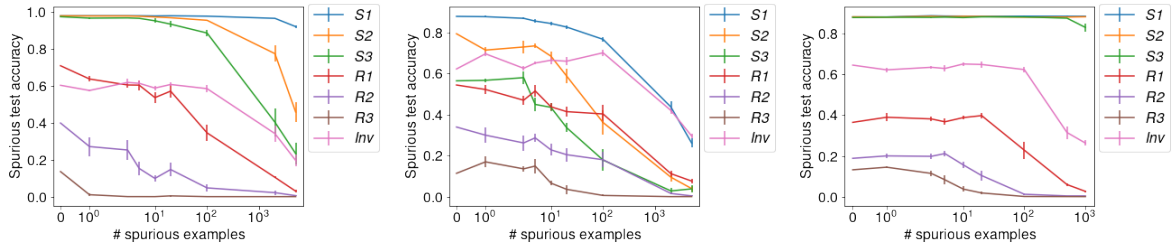
Moreover, in Appendix B, we explore rare spurious correlation theoretically. The theoretical findings aligns with our observations here. Inspired by our theoretical results, we also explore ways to mitigate these rare spurious correlations in Appendix C. Finally, related works are in Appendix D.

Discussion. Our experimental results suggest that neural networks are *highly* sensitive to very small amounts of spurious training data. Furthermore, the learnt rare spurious correlations cause undesirable effects on privacy and test accuracy. Easy learning of rare spurious correlations can lead to privacy issues (Leino & Fredrikson, 2020) – where an adversary may infer the presence of a confidential image in a training dataset based on output probabilities. It also raises fairness concerns as a neural network can draw spurious conclusions about a minority group if a small number of subjects from this group are present in the training set (Izzo et al., 2021). We recommend to test and audit neural networks thoroughly before deployment in these applications.



(a) MNIST, $c_{\text{tar}} = 0$, *R3* (b) Fashion, $c_{\text{tar}} = 0$ (c) CIFAR10, $c_{\text{tar}} = 0$, *R3*

Figure 3. The test accuracy of the membership inference attack model on all examples vs. spurious examples.



(a) MNIST, $c_{\text{tar}} = 0$ (b) Fashion, $c_{\text{tar}} = 0$ (c) CIFAR10, $c_{\text{tar}} = 0$

Figure 4. Each figure shows the mean and standard error of the spurious scores on three datasets, MNIST, Fashion and CIFAR10, and different numbers of spurious examples.

3. Conclusion

We demonstrate that rare spurious correlations are learnt readily by neural networks, and we look closely into this phenomenon. We discover that a few spurious examples can lead to the model learning the spurious correlation through empirical and theoretical evidence. We then show that these rare spurious correlations can have impact on both privacy and test accuracy of the model. Finally, we find that some regularization methods, including weight decay and adding Gaussian noise to the input, can reduce the spurious correlation without sacrificing accuracy both empirically and theoretically. However, they are far from completely removing these spurious correlations.

Acknowledgements

We thank Angel Hsing-Chi Hwang for providing thoughtful comments on the paper. This work was supported by NSF under CNS 1804829 and ARO MURI W911NF2110317. CNC's research is supported by Simons Investigator Fellowship, NSF grants DMS-2134157 and CCF-1565264,and DOE grant DE-SC0022199.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

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

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pp. 233–242, 2017.

Bahng, H., Chun, S., Yun, S., Choo, J., and Oh, S. J. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539, 2020.

Basu, S., Pope, P., and Feizi, S. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020a.

Basu, S., You, X., and Feizi, S. On second-order group influence functions for black-box predictions. In *Inter-*

national Conference on Machine Learning, pp. 715–724, 2020b.

Burkard, C. and Lagesse, B. Analysis of causative attacks against svms learning from data streams. In *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*, pp. 31–36, 2017.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 267–284, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.

Chang, C.-H., Adam, G. A., and Goldenberg, A. Towards robust classification model by counterfactual and invariant data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15212–15221, 2021.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284, 2006.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. *arXiv preprint arXiv:2008.03703*, 2020.

Ginart, A., Guan, M. Y., Valiant, G., and Zou, J. Making ai forget you: Data deletion in machine learning. *arXiv preprint arXiv:1907.05012*, 2019.

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

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Hartley, J. and Tsaftaris, S. A. Measuring unintended memorisation of unique private features in neural networks. *arXiv preprint arXiv:2202.08099*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016, 2021.

Khani, F. and Liang, P. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 196–205, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894. PMLR, 2017.

Koh, P. W., Ang, K.-S., Teo, H. H., and Liang, P. On the accuracy of influence functions for measuring group effects. *arXiv preprint arXiv:1905.13289*, 2019.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images, 2009.

Kulynych, B., Yang, Y.-Y., Yu, Y., Błasiok, J., and Nakkiran, P. What you see is what you get: Distributional generalization for algorithm design in deep learning. *arXiv preprint arXiv:2204.03230*, 2022.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Leino, K. and Fredrikson, M. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pp. 1605–1622, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

Min, Y., Chen, L., and Karbasi, A. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *arXiv preprint arXiv:2002.11080*, 2020.

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

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning. *arXiv preprint arXiv:2007.02923*, 2020.

pandas development team, T. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

Sablayrolles, A., Douze, M., Schmid, C., and Jégou, H. Radioactive data: tracing through training. In *International Conference on Machine Learning*, pp. 8326–8335, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

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

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

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

Srivastava, M., Hashimoto, T., and Liang, P. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pp. 9109–9119, 2020.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wang, Y. and Chaudhuri, K. Data poisoning attacks against online learning. *arXiv preprint arXiv:1808.08994*, 2018.

Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C., and Roli, F. Support vector machines under adversarial label contamination. *Neurocomputing*, 160:53–62, 2015.

Xiao, K., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Yeom, S., Fredrikson, M., and Jha, S. The unintended consequences of overfitting: Training data inference attacks. *arXiv preprint arXiv:1709.01604*, 12, 2017.

Zhou, C., Ma, X., Michel, P., and Neubig, G. Examining and combating spurious features under distribution shift. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12857–12867, Jul 2021.