Monitoring Shortcut Learning using Mutual Information

Mohammed Adnan [1 2] **Yani Ioannou** [3] **Chuan-Yung Tsai** [2] **Angus Galloway** [1 2]
H.R. Tizhoosh [4 5] **Graham W. Taylor** [1 2]

Abstract

The failure of deep neural networks to generalize to out-of-distribution data is a well-known problem and raises concerns about the deployment of trained networks in safety-critical domains such as healthcare, finance and autonomous vehicles. We study a particular kind of distribution shift — *shortcuts* or *spurious correlations* in the training data. Shortcut learning is often only exposed when models are evaluated on real-world data that does not contain the same spurious correlations, posing a serious dilemma for AI practitioners to properly assess the effectiveness of a trained model for real-world applications. In this work, we propose to use the mutual information (MI) between the learned representation and the input as a metric to find where in training the network latches onto shortcuts. Experiments demonstrate that MI can be used as a domain-agnostic metric for monitoring shortcut learning.

1. Introduction

Our understanding of 'how' and 'what' neural networks learn is limited, which raises concern about the deployment of neural networks in safety-critical domains. Despite achieving state-of-the-art performance on benchmark datasets, neural networks may fail to generalize in real-world settings or for out-of-distribution data (Koh et al., 2021). For example, models trained for cancer detection may not generalize on data from a new hospital (Castro et al., 2020; Perone et al., 2018; AlBadawy et al., 2018) and self-driving cars may not generalize to new lighting conditions or object poses (Alcorn et al., 2018; Dai & Van Gool, 2018). One reason why models may fail in real-world settings could be attributed to learning *shortcuts* (Geirhos et al., 2020) from the training data. A *shortcut* is a type of distribution shift where spurious correlations exist only in the training data, resulting in the learning of non-intended or easy-to-learn discriminatory features which work well on the training and test dataset but not on out-of-distribution real-world datasets (Wiles et al., 2022; Geirhos et al., 2020). Shortcuts can arise due to dataset biases or the model using 'trivial' or unintended features like high-frequency noise patterns or the image background for the classification task. For example, an Inception-V3 model trained to detect hip fractures was found to use scanner information for learning discriminatory features (Badgeley et al., 2019); deep learning systems trained to detect COVID-19 from chest radiographs can rely on confounding factors (shortcuts) rather than medical pathology (DeGrave et al., 2021). While our understanding of shortcuts and how they arise is still developing, a helpful tool to practitioners deploying machine learning models in safety-critical domains with a high cost of failure would be to monitor shortcuts during the training phase. Although the phenomenon of shortcut learning is widely known, there is no effective method available to monitor shortcuts being learned. Interpretable machine-learning methods such as feature attribution, Grad-CAM (Selvaraju et al., 2017), and LIME (Ribeiro et al., 2016) have been used to understand a model's dependency on spurious correlations. However, it has been shown that such post-hoc explanations are ineffective (Adebayo et al., 2021; Alqaraawi et al., 2020; Chu et al., 2020).

In this work, we show that shortcut learning can be understood using the information bottleneck framework (Tishby et al., 1999; Tishby & Zaslavsky, 2015), by using the mutual information (MI) between the inputs and the learned representation to monitor shortcut learning. We use the neural tangent kernel (NTK) (Jacot et al., 2018) to study the training evolution of shortcut learning. We design experiments using synthetic and complex real-world data to demonstrate the relationship between (MI) and shortcut learning, and show that MI can be used as a metric for domain-agnostic assessment of shortcuts. We find that compression as measured by MI is associated with the tendency to learn shortcuts.

[1]University of Guelph [2]Vector Institute, Canada [3]University of Calgary [4]Mayo Clinic, Rochester, MN, USA [5]Kimia Lab, University of Waterloo, Canada. Correspondence to: Mohammed Adnan <madnan01@uoguelph.ca>, Graham Taylor <gwtaylor@uoguelph.ca>.

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

2. Background

Shortcut learning: Wiles et al. (2022) defined shortcuts or spurious correlations as a type of *distribution shift* such that two or more attributes are correlated at training time, but not for the test data, where they are independent. In a more general sense, shortcuts are *easy-to-learn* decision rules that can be exploited in the absence of distribution shift (i.e. on standard benchmarks) but fail to transfer to more challenging and diverse testing conditions, such as real-world datasets (Geirhos et al., 2020).

Information bottleneck method: The information bottleneck (IB) can be viewed as a rate-distortion problem cast entirely in terms of mutual information (MI)—denoted "$I(V;Z) = I(Z;V)$" (Tishby et al., 1999). A *distortion* function measures how well a relevant variable V is predicted from another variable Z, where Z is usually a compressed representation of the input X. The *rate* refers to the complexity of Z, which is less than or equal to X. IB is a general method for data compression but has been advanced as an explanatory tool predictive of learning and generalization of neural networks (NNs). It has been suggested that NNs trained by SGD may learn compressed representations Z of their input, making them insensitive to data idiosyncrasies, yet maintain sufficient *relevant* information for predicting the output Y (e.g. class labels) (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017). This trade-off between compression and preserving task relevant information is optimized by the notion of "minimal sufficient statistics" (Cover & Thomas, 1991). The IB view suggests that NNs trained by cross-entropy loss may implicitly minimize the following Lagrangian:

$$\min I(X;Z) - \beta I(Z;Y), \tag{1}$$

enabling them to implement minimal sufficient statistics for different β-constraints on the error.[1]

Neural tangent kernel: Distribution-free estimation of MI for high-dimensional data is challenging and often intractable. One workaround to this problem is using an infinite ensemble of infinite-width neural networks (Shwartz-Ziv & Alemi, 2020) which confer tractable bounds on MI. The neural tangent kernel (NTK) (Jacot et al., 2018) is a kernel that describes the evolution of infinite-width neural networks during their training by gradient descent, thus allowing the systematic study of neural networks using tools from kernel methods. Infinite-width neural networks behave as linear functions, and their training evolution can be fully described by the NTK. Shwartz-Ziv & Alemi (2020) used the fact that the output of an infinite ensemble of infinitely-wide neural networks initialized with Gaussian weights and biases and trained by MSE loss is a conditional Gaussian distribution. This allows tractable computation of (MI) between the representation Z and the targets Y: $I(Z;Y)$, and the MI between Z and the inputs X during training: $I(X;Z)$.

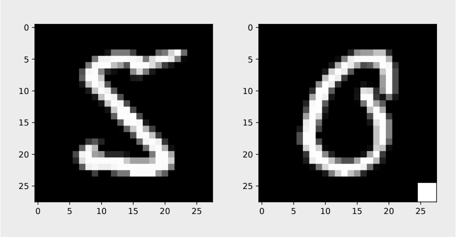


Figure 1. Sample images from MNIST dataset. A small patch is added to all images of even digits as a shortcut.

3. Shortcuts and information bottleneck

Our hypothesis is that MI as measured during the evolution of a network's parameters can be used to monitor shortcut learning as it occurs. Shortcuts allow networks to learn a maximally compressed representation Z, i.e., $I(X;Z)$ is considerably reduced and thus can be used as a metric to monitor exploitation of shortcuts.

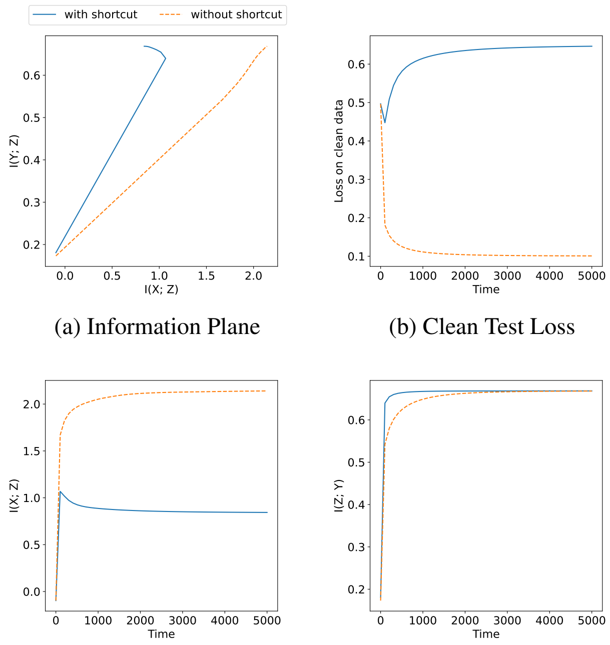
To support our hypothesis, we design controlled datasets containing spurious correlations and measure $I(X;Z)$ during the training evolution using the NTK. Using insights from the concept of a previously introduced "mutual information plane" (Shwartz-Ziv & Tishby, 2017), we can observe shortcut learning as it happens and find the time where the network stops exploring the preferred (generalizable) solution space and latches onto the spurious signal.

4. Experiments and observations

Experimental Setup: For each experiment, we plot $I(X;Z)$ and $I(Z;Y)$ w.r.t. time, generalization error/loss on clean data (without spurious correlations), and the information plane ($I(Z;Y)$ vs. $I(X;Z)$). Note that we calculate the upper bound of $I(X;Z)$, and are only interested in the training dynamics rather than the precise value of MI.

MNIST with synthetic shortcut: We train a model to classify MNIST images into odd and even digits. We add a small white patch on one corner of *all even* digits of the MNIST training dataset as a spurious correlation (Figure 1). The network can use the patch to accurately classify the images into odd and even. We compare the MI during the training evolution on datasets with and without shortcuts.

[1] To what extent the relationship between IB and deep learning holds in general is the subject of ongoing debate (Saxe et al., 2018; Jacobsen et al., 2018; Goldfeld et al., 2019).



(a) Information Plane

(b) Clean Test Loss

(c) $I(X;Z)$

(d) $I(Z;Y)$

Figure 2. Comparison of training evolution with and without synthetic shortcuts on MNIST dataset. An upper bound on MI is plotted. (a) Information plane plotting $I(X;Z)$ versus $I(Z;Y)$ during training. (b) Loss w.r.t. the clean test data without shortcuts—the solid blue line corresponds to a model trained with shortcuts, the broken orange line is without shortcuts. (c) Plot of $I(X;Z)$ versus training time step. (d) Plot of $I(Y;Z)$ during the training evolution. Animated GIF of the plot can be viewed here.

In Figure 2a and Figure 2c, we observe that the mutual information $I(X;Z)$ increases initially during training but then latches onto the shortcuts, after which the mutual information decreases sharply. It can also be observed in Figure 2b that generalization error increases after the point at which MI starts to decrease, indicating that the network explores the more optimal region of the solution space in the initial training epochs before discovering shortcuts. This is consistent with the findings of Schwartz-Ziv and Tishby (2017) about the behaviour of SGD. In the initial phase, SGD explores the multidimensional space of solutions. When it begins converging, it arrives at the diffusion phase in which the network learns to compress (Shwartz-Ziv & Tishby, 2017). In both settings, the model achieves high training accuracy, i.e., $I(Z;Y)$ (Figure 2d) but the difference in test set loss is striking (Figure 2b).

Visualisation: To visually verify that the network is learning the shortcut in the above experiment, we generate a saliency map. We use the finite-difference estimation to find the gradient of the class probability w.r.t. to the input pixels. The network predominantly uses the shortcut in the image to predict the class (Figure 3).

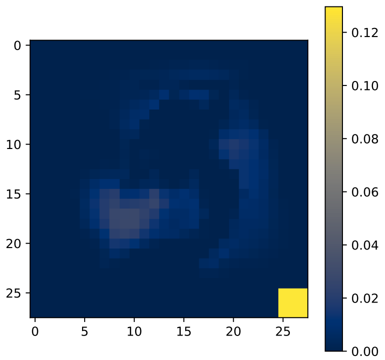


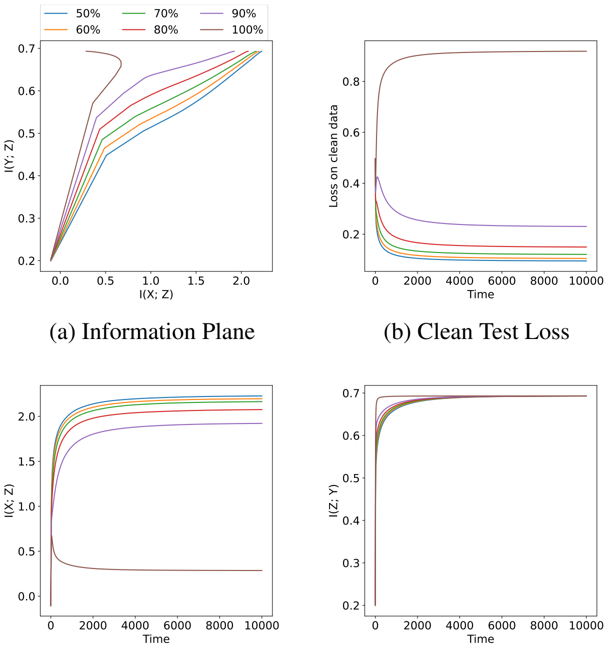
Figure 3. Saliency map of model trained on MNIST images with a small patch on even images as a shortcut.

Effect of partially correlated shortcuts: In real-world data, shortcuts are often partially correlated with the output, i.e., the model cannot classify with 100% accuracy using only the shortcuts. To understand the effect of partially correlated shortcuts on the training dynamics, we construct different datasets with varying degrees of shortcut efficacy. Instead of corrupting all the even images, we add a small white patch on one corner only to a specific percentage (50%–100%) of even images. We plot the MI for varying degrees of corruption in Figure 4 for 1000 training points sampled uniformly.

We observe that as the effectiveness of the shortcut increases, $I(X;Z)$ converges to a lower value indicating the ability to perform more compression. We also note an interesting behaviour during the training evolution: when the shortcut is partially correlated, MI does not decrease significantly as compared to the 100% effective shortcut. We speculate that, while in these cases (e.g. 80–90% shortcut efficacy) the model is able to recover some generalization ability, its ability to discover high generalization solutions is irrevocably deteriorated once it discovers the minima corresponding to the shortcut solution.

CelebA with natural shortcuts: We also test our hypothesis on a dataset containing natural shortcuts. We curate images from the CelebA dataset such that all images tagged as male have the “black hair color” attribute, while images tagged as female have the “blonde hair color” attribute. We train the network to classify facial images into the male and female categories[2], while the network may use hair color as a shortcut for accurately classifying the images. Since the dataset is not controlled, some images from both classes have a black color in different parts of the image (background, clothes, etc.), reducing the effectiveness of the hair color attribute. We plot the MI trajectory with and without

[2] The CelebA dataset only provides binary labels and we do not know how the gender attribute was assigned. Therefore it should be considered as nothing more than an arbitrary class in this experiment.



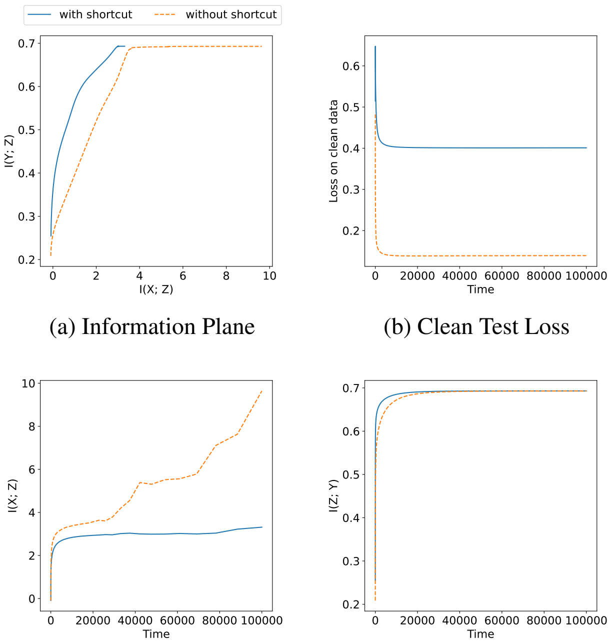
(a) Information Plane (b) Clean Test Loss

(c) $I(X;Z)$ (d) $I(Z;Y)$

Figure 4. Effect of shortcut effectiveness on the MI trajectory. Shortcut is added to different percentages of even images in each experiment. With increase in shortcut effectiveness, MI converges to a lower value. Animated GIF of the plot can be viewed here.

shortcuts for 100 sample training points sampled uniformly in log scale (Figure 5). We observe MI profile similar to Figure 4. On data without the shortcut, MI increases consistently, while in the presence of shortcuts, MI converges to a lower value, therefore validating our hypothesis on real-world data with natural shortcuts.

Effect of shortcut on the loss landscape: We visualize the loss landscape of neural networks to understand the effect of shortcuts on the optimization trajectory. We plot loss along a linear path connecting the initial parameter θ_o and converged parameter θ^* in the weight space (Goodfellow et al., 2014) and polar coordinates (r_t, ϕ_t) plot measuring the deviation from the linear line between θ_i and θ^* (Figure 6). We parameterize the line with α such that $\theta = (1-\alpha)\theta_i + \alpha\theta^*$. Polar coordinates can be calculated using $r_t = \frac{||\triangle\theta_t||}{||\triangle\theta_o||}$ and $\phi_t = \arccos\frac{\triangle\theta_t \times \triangle\theta_o}{||\triangle\theta_t|| \times ||\triangle\theta_o||}$, where $\triangle\theta_t = \theta_t - \theta^*$. We observe that the loss landscape around θ^* in the case of shortcuts is surprisingly flat as compared to the valley-like shape for a model trained on data not containing shortcuts using the MNIST dataset. The polar plot shows that the optimizer deviates less from the linear trajectory when trained with shortcuts.



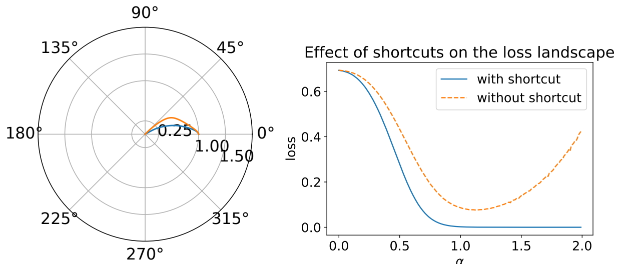
(a) Information Plane (b) Clean Test Loss

(c) $I(X;Z)$ (d) $I(Z;Y)$

Figure 5. Mutual information profile for CelebA dataset with hair attribute as shortcut. Plot of $I(X;Z)$ in (a) and (c) show that shortcuts results in reduced $I(X;Z)$. Animated GIF of the plot can be viewed here.

5. Conclusion and future work

In this work, we sought to understand why networks tend to to learn shortcuts through the lens of the information bottleneck method. We showed that mutual information can be used to monitor training dynamics w.r.t. shortcut learning without using any domain knowledge; this is an advantage compared to methods adapted by the interpretable ML literature, where domain knowledge is required. However, we used the NTK to estimate mutual information, limiting our



(a) Radial plot (b) Loss landscape visualization

Figure 6. Visualization of loss landscape on MNIST dataset. (a). polar coordinates (r_t, ϕ_t) measuring the deviation from the linear path between initialisation and converged parameters in the weight space during the optimization. (b). 1-D visualization of loss landscape.

approach to infinite-width neural networks. In the future, we will address this limitation and develop methods to avoid shortcut learning using tractable MI estimators that place less severe constraints on the model architecture (Selby et al., 2022; Gabrié et al., 2018). We are also interested in exploring the relationship between the curvature of the solution minima and shortcut learning.

References

Adebayo, J., Muelly, M., Abelson, H., and Kim, B. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International Conference on Learning Representations*, 2021.

AlBadawy, E. A., Saha, A., and Mazurowski, M. A. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3):1150–1158, 2018.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects, 2018.

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Berthouze, N. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 275–285, 2020.

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B. L., Snyder, T. M., and Dudley, J. T. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2, 2019.

Castro, D. C., Walker, I., and Glocker, B. Causality matters in medical imaging. *Nature Communications*, 11, 2020.

Chu, E., Roy, D., and Andreas, J. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*, 2020.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

Dai, D. and Van Gool, L. Dark model adaptation: Semantic image segmentation from daytime to nighttime, 2018.

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

Gabrié, M., Manoel, A., Luneau, C., Macris, N., Krzakala, F., Zdeborová, L., et al. Entropy and mutual information in models of deep neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. April 2020.

Goldfeld, Z., Van Den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., and Polyanskiy, Y. Estimating Information Flow in Deep Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2299–2308. PMLR, 2019.

Goodfellow, I. J., Vinyals, O., and Saxe, A. M. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014.

Jacobsen, J.-H., Smeulders, A. W. M., and Oyallon, E. I-RevNet: Deep Invertible Networks. In *International Conference on Learning Representations*, 2018.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. June 2018.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Perone, C. S., Ballester, P., Barros, R. C., and Cohen-Adad, J. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. ”why should i trust you?”: Explaining the predictions of any classifier, 2016.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. On the Information Bottleneck Theory of Deep Learning. In *International Conference on Learning Representations*, 2018.

Selby, K. A., Rashid, A., Kobyzev, I., Rezagholizadeh, M., and Poupart, P. Learning functions on multiple sets using multi-set transformers. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. URL `https://openreview.net/forum?id=HzMEEOUs5x5`.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shwartz-Ziv, R. and Alemi, A. A. Information in infinite ensembles of infinitely-wide neural networks. In Zhang, C., Ruiz, F., Bui, T., Dieng, A. B., and Liang, D. (eds.), *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pp. 1–17. PMLR, 08 Dec 2020. URL https://proceedings.mlr.press/v118/shwartz-ziv20a.html.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. 2017. *arXiv preprint arXiv:1703.00810*, 2017.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Allerton Conference on Communication, Control and Computing*, 1999.

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S.-A., Ktena, I., Dvijotham, K. D., and Cemgil, A. T. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022.