Measures of Information Reflect Memorization Patterns

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural networks are known to exploit spurious artifacts (or shortcuts) that co-occur with a target label, exhibiting *heuristic memorization*. On the other hand, networks have been shown to memorize training examples, resulting in *example-level memorization*. These kinds of memorization impede generalization of networks beyond their training distributions. Detecting such memorization could be challenging, often requiring researchers to curate tailored test sets. In this work, we hypothesize—and subsequently show—that the diversity in the activation patterns of different neurons is reflective of model generalization and memorization. We quantify the diversity in the neural activations through information-theoretic measures and find support for our hypothesis on experiments spanning several natural language and vision tasks. Importantly, we discover that information organization points to the two forms of memorization, even for neural activations computed on unlabeled in-distribution examples. Lastly, we demonstrate the utility of our findings for the problem of model selection.

1 Introduction

Current day deep learning networks are limited in their ability to generalize across different domains and settings. Prior studies found that these networks rely on spurious artifacts that are correlated with a target label (Schölkopf et al., 2012; Lapuschkin et al., 2019; Geirhos et al., 2019, 2020, inter alia). We refer to learning of such artifacts (also known as heuristics or shortcuts) as *heuristic memorization*. Further, neural networks can also memorize individual training examples and their labels; for instance, when a subset of the examples are incorrectly labeled (Zhang et al., 2017; Arpit et al., 2017; Tänzer et al., 2021). We refer to this behavior as *example-level memorization*. A large body of past work has established that these facets of memorization pose a threat to generalization, especially on out-of-distribution (OOD) scenarios where the memorized input features and corresponding target mappings do not hold (Ben-David et al., 2010; Wang et al., 2021b; Hendrycks et al., 2021a; Shen et al., 2021). To stimulate such OOD distributions, however, researchers are required to laboriously construct specialized labeled examples to measure the extent of suspected fallacies in models. While these sets make it possible to assess model behavior over a chosen set of features, the larger remaining features remain hard to identify and study. Performance measures to assess models over these sets, moreover, are not indicative of internal workings that manifest them, limiting their utility in improving models.

In this work, we posit that organization of information across internal activations of a network could be indicative of memorization. Consider a sample task of separating concentric circles, illustrated in Figure 1a. It is straightforward for a two-layered feed-forward network to identify the circular decision boundary for this task. However, if the nature of this learning task is changed, the network

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

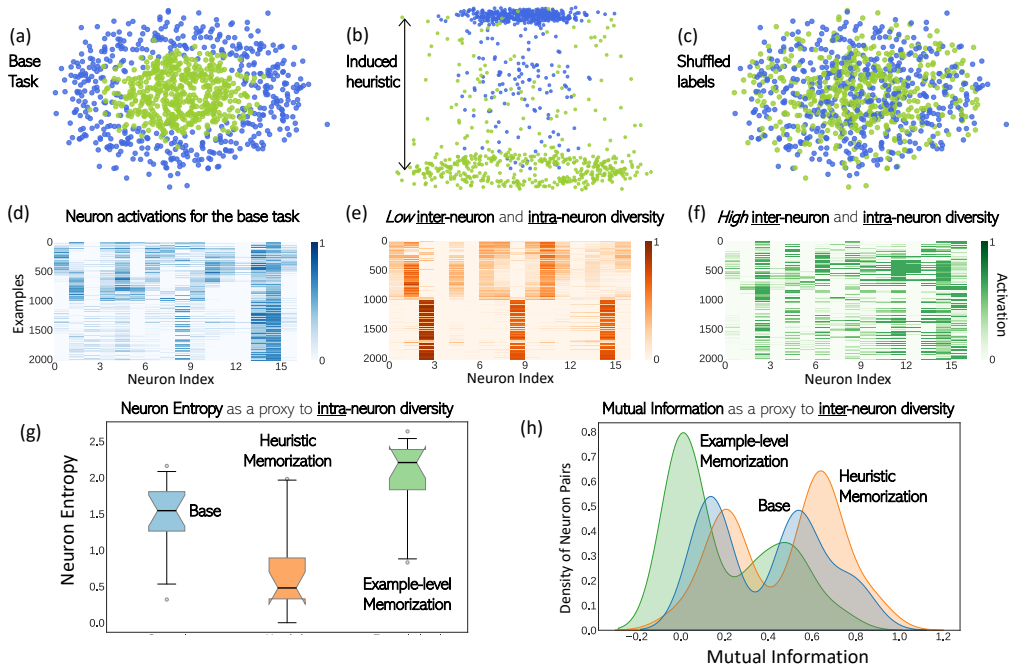

Figure 1: (**a**) A toy setup of separating concentric circles; (**b**) An additional feature spuriously simplifies the task, inciting *heuristic memorization*; (**c**) Shuffled target labels induce *example-level memorization*; (**d**) Neuron activations for a two-layered feed-forward network trained for the base task in (a); (**e**) Activation patterns for the network reflect low *intra-neuron* and *inter-neuron* diversity when trained on (b); (**f**) High *intra-neuron* and *inter-neuron* diversity is seen when the network is trained on (c); (**g**) *Entropy* acts as a proxy to intra-neuron diversity; (**h**) *Mutual Information* acts as a proxy to inter-neuron diversity. Distinguishable patterns for the three networks are seen in (g) and (h).

may resort to memorization. When a spurious feature is introduced in this dataset such that its value ($+/-$) correlates to the label (0/1) (Figure 1b), the network *memorizes* the feature-to-label mapping, reflected in a uniform activation pattern across neurons (Figure 1e). In contrast, when labels for the original set are shuffled (Figure 1c), the same network memorizes individual examples during training and shows a high amount of diversity in its activation patterns (Figure 1f). This example demonstrates how memorizing behavior is observed through diversity in neuron activations.

We formalize the notion of *diversity* across neuron activations through two measures: (i) *intra-neuron* diversity: the variation of activations for a neuron across a set of examples, and (ii) *inter-neuron* diversity: the dissimilarity between pairwise neuron activations on the same set of examples. We hypothesize that the nature of these quantities for two networks could point to underlying differences in their generalizing behavior. In order to quantify intra-neuron and inter-neuron diversity, we adopt the information-theoretic measures of *entropy* and *mutual information* (MI), respectively.

Throughout this work, we investigate if diversity across neural activations (§2) reflects model generalizability. We compare networks with varying levels of heuristic (§3) or example-level (§4) memorization across a variety of settings: synthetic setups based on the `IMDb` (Maas et al., 2011) and `MNIST` (Lecun et al., 1998) datasets for both memorization types, as well as naturally occurring scenarios of gender bias on `Bias-in-Bios` (De-Arteaga et al., 2019) and OOD image classification on `NICO` (Zhang et al., 2022). We find that the information measures consistently capture differences among networks with varying degrees of memorization: Low entropy and high MI are characteristic of networks that show heuristic memorization, while high entropy and low MI are indicative of example-level memorization. Lastly, we evaluate these measures from the viewpoint of model selection and note strong correlations to rankings from domain-specific evaluation metrics (§5).

2 Methods

As per the data processing inequality (Beaudry & Renner, 2012), a part of the neural network (referred to as the *encoder*) compresses the most relevant information of a given input X, into a representation H. This compressed information is processed by a *classification head* (or, the *decoder*) to produce an output Y corresponding to the given input. We hypothesize that the organization of information across neurons of the encoder is indicative of model generalization. We study two complementary properties that, in our view, capture the information organization:

(i) **Intra-neuron diversity**: How do activations of a given neuron vary across different input examples. We measure the *entropy* of neural activations (across examples) as a proxy.

(ii) **Inter-neuron diversity**: How unique is the activation of a neuron compared to other neurons. We quantify this via *mutual information* between activations of pairwise neurons.

Below, we discuss the information measures formally.

2.1 Information Measures

For any given encoder (consisting of N neurons) that maps the input to a dense hidden representation, we denote the activation of the ith neuron as a random variable, $A_i \in \{a_i^1, \ldots, a_i^S\}$, where each measurement is an activation over an example from a set of size S.[1] The probability over this continuous activation space is computed by binning it into discrete ranges (Darbellay & Vajda, 1999), and we denote each discretized activation value as $\hat{a}$.

Entropy We measure the Shannon entropy for each neuron in the concerned network, as a proxy of intra-neuron diversity. Following the definition of Shannon entropy, this is given as:

$$H(A_i) = \mathop{\mathbb{E}}_{\hat{a}_i^s \in A_i} [h(\hat{a}_i^s)] = \frac{1}{S} \sum_{s=1}^{S} p(\hat{a}_i^s) \log(\frac{1}{p(\hat{a}_i^s)}) \tag{1}$$

Mutual Information We compute the mutual information (MI) between underlying neurons as a proxy to inter-neuron diversity. Specifically, we compute the MI between all neuron pairs in the network.[2] Thus, the set of MI values $I(A_i)$ for a particular neuron A_i, is given as:

$$I(A_i) = \{I(A_i; A_1), \ldots, I(A_i; A_N)\} \tag{2}$$

where, $I(X;Y)$ depicts the MI between variables X and Y. Unless stated otherwise, this $I(A_i)$ is computed $\forall\, i \in \{1, \ldots, N\}$, resulting into a square matrix of size $(N \times N)$.

This process of computing the information measures for a network on a given set of examples is summarized in Algorithm 1. Further details on the computation are given in appendix A.

Algorithm 1 Computation of information measures. Algorithmic procedures ENTROPY and MI are given in algorithms 2 and 3, respectively, in appendix A.

```
A_1, ..., A_N ← {f(x_i)}_{i=1}^S
▷ Computing activations for all neurons
H ← {}; I ← {}
for i ∈ {1, ..., N} do
    I_i ← {}
    H_i ← ENTROPY(A_i)                    ▷ Eq. 1
    for j ∈ {1, ..., N} do
        I_i ← I_i ⊕ MI(A_i, A_j)          ▷ Eq. 2
    end for
    H ← H ⊕ H_i
    I ← I ⊕ I_i
end for
```

[1] In this work, this set of examples resembles the input distribution on which the underlying network is trained.

[2] In principle, we would compute MI across neuron sets; we approximate this through individual neuron pairs.

2.2 Toy setup: concentric circles

Here, we briefly discuss the information-theoretic metrics for the example of concentric circles from the introduction (Figure 1). To recap, we consider a setup to compare networks showing the two forms of memorization and observe discernible differences in their activation patterns: heuristic memorization corresponds to low intra-neuron and inter-neuron diversity, while example-level memorization corresponds to high diversity (Figures 1e and 1f). We expect that this difference in diversity would be captured through the above defined information measures.

Figure 1g presents the distribution of entropy values for each of the three networks with varying generalization behaviors. Throughout this work, we visualize this distribution of entropy using similar box-plots, where a black marker within the boxes depicts the median of the distribution and a notch neighboring this marker depicts the 95% confidence interval around the median. We observe that entropy for the network exhibiting heuristic memorization is distributed around a lower point than the others, whereas entropy for the network with example-level memorization is higher.

Furthermore, Figure 1h shows the distribution of MI for the three networks. To interpret the distribution of MI (an $N \times N$ square matrix), we fit a Gaussian mixture model over all values and visualize it through a density plot, where the density (*y-axis*) at each point corresponds to the number of neurons pairs that exhibit that MI value (*x-axis*). Larger peaks in these density plots suggest a large number of neurons pairs are concentrated in that region. Interestingly, we see such peaks for the three networks at distinct values of MI. For the network showing example-level memorization (high inter-neuron diversity), most of the neuron pairs show low values of MI. In contrast, heuristic memorization (low inter-neuron diversity) has high neuron pair density for higher MI values.[3]

Based on these findings, we formulate two hypotheses, summarized in Table 1:

H1 Networks exhibiting heuristic memorization would show low inter- and intra-neuron diversity, reflected through low entropy and high MI values.

H2 Networks exhibiting example-level memorization would show high inter- and intra-neuron diversity, reflected through high entropy and low MI values.

Table 1: Summarizing our hypotheses.

Memorization	Diversity	
	Intra-neuron ($\propto$ Entropy)	Inter-neuron ($\propto$ MI^{-1})
Heuristic	$\downarrow$	$\downarrow$
Example-level	$\uparrow$	$\uparrow$

3 Heuristic Memorization

Here, we study different networks with varying degrees of heuristic memorization, and examine if the information measures—aimed to capture neuron diversity—indicate the extent of memorization.

3.1 Semi-synthetic Setups

We synthetically introduce spurious artifacts in the training examples such that they co-occur with target labels. Networks trained on such a set are prone to memorizing these artifacts. The same correlations with an artifact do not hold in the validation sets. To obtain a set of networks with varying degrees of this heuristic memorization, we consider a parameter α that controls the fraction of the training examples for which the spurious correlation holds true. We consider the following setups:

Colored MNIST In this setting, the MNIST dataset (Lecun et al., 1998) is configured such that a network trained on this set simply learns to identify the color of images and not the digits themselves (Arjovsky et al., 2019). Particularly, digits 0–4 are grouped as one label while 5–9 as the other, and images for these labels are colored green and red, respectively. For this setup, we train multi-layer perceptron (MLP) networks for varying values of α, which corresponds to the fraction of training instances that abide to the color-to-label correlation.

[3]This difference in neuron activation patterns for the two memorizing sets could be caused by several factors, including functional complexity (Lee et al., 2020): Functions that encode individual data points (as in example-level memorization) need to be much more complex than functions that learn shortcuts (heuristic memorization).

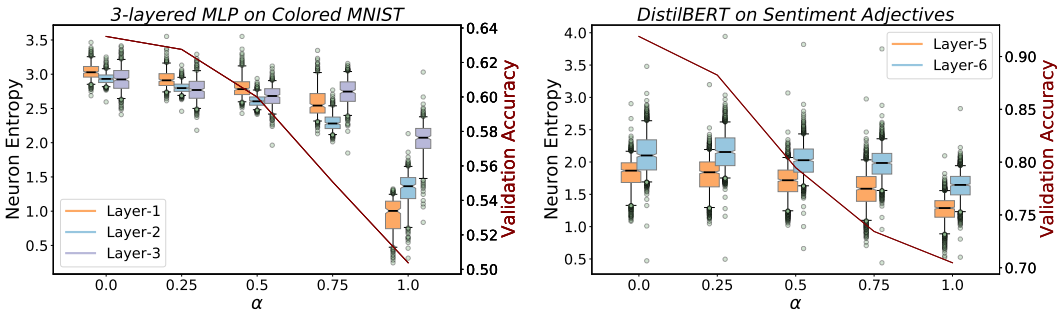

Figure 2: The relation between entropy of neural activations and heuristic memorization. For both the setups, networks trained on higher α show higher heuristic memorization (as depicted by the dipping model accuracy line), accompanied with lower entropy values.

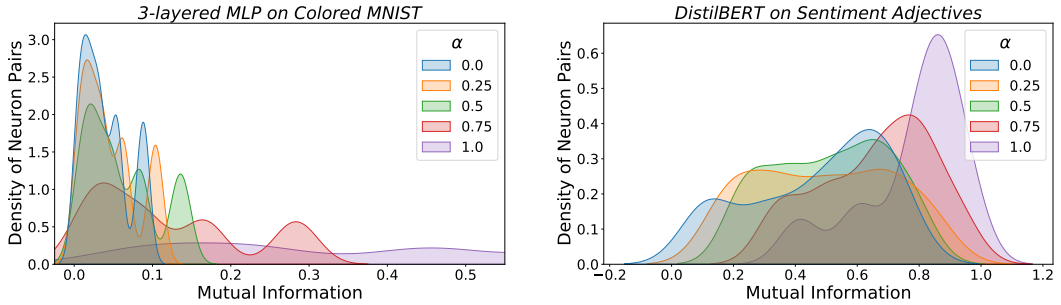

Figure 3: Distribution of mutual information (MI) of pairs of neurons for networks with varying heuristic memorization. For both settings, networks trained on training sets with larger amounts of spurious correlations ($\uparrow \alpha$) exhibit higher mutual information across their neuron pairs.

Sentiment Adjectives In this setup, we sub-sample examples from the `IMDb` dataset (Maas et al., 2011) that contain at least one adjective from a list of positive and negative adjectives. Then, examples that contain any of the positive adjectives ("good", "great", etc.) are marked with the positive label, whereas ones that contain any negative adjectives ("bad", "awful", etc.) are labeled as negative. We exclude examples that contain adjectives from both lists. The motivation to use this setup is to introduce heuristics in the form of adjectives in the training set. We fine-tune DistilBERT-base models (Sanh et al., 2019) on this task for different values of α (fraction of examples that obey the heuristic).[4]

Results: Through these experiments, we first note that **low entropy across neural activations indicates heuristic memorization in networks**. This is evident from Figure 2, where we see that (1) as we increase α the validation performance decreases, indicating heuristic memorization (see the solid line in the plots); and (2) with an increase in this heuristic memorization, we see lower entropy across neural activations. We show the entropy values of neural activations for the 3 layers of an MLP trained on `Colored MNIST` (left sub-plot) and for the last two layers of DistilBERT on `Sentiment Adjectives` (right sub-plot)[5]. In both these two scenarios, we see a consistent drop in the entropy with increasing values of α, with a particularly sharp decline when $\alpha = 1.0$.

Furthermore, we observe that **networks with higher heuristic memorization exhibit higher mutual information** across pairs of neurons. In Figure 3, networks with higher memorization ($\uparrow \alpha$), have larger density of neurons in the high mutual information region. While this trend is consistent across the two settings, we see some qualitative differences: The memorizing ($\alpha = 1.0$) MLP network

[4]More details regarding these datasets and training procedures are given in appendices B.1 and B.2.

[5]Considerable changes in entropy values are not seen for initial DistilBERT layers, suggesting that spurious correlations are largely captured by later layers. Detailed results covering other layers are given in appendix C.1.

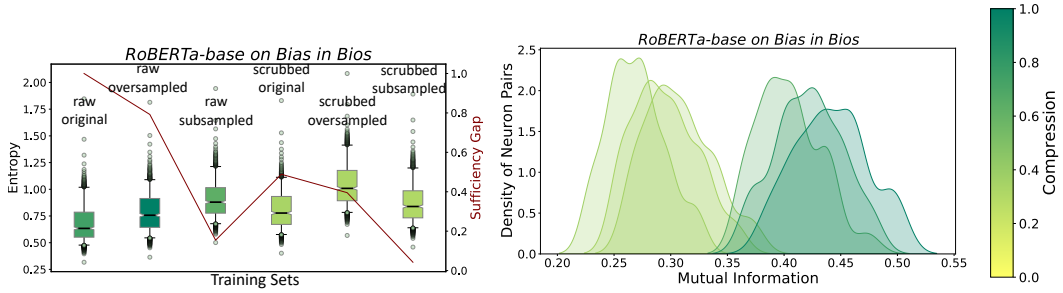

Figure 4: Distributions of entropy and MI across final layer activations of RoBERTa-base differentiate networks fine-tuned on original and de-biasing sets for `Bias-in-Bios`. Color of boxes and Gaussian plots corresponds to *extractability* of gender information in model representations as estimated through MDL probing(Voita & Titov, 2020) —lighter colors indicate lower extractability (less bias).

on `Colored MNIST` (left) has a uniform distribution across the entire scale of MI values, while DistilBERT on `Sentiment Adjectives` (right) largely has a high-density peak for an MI of ~ 0.9.

3.2 Natural Setups

Next, we investigate setups where spurious correlations are not synthetically induced, but occur naturally in the datasets. Below, we describe two such scenarios:

Occupation Prediction We first study the task of predicting occupations from biographies on the `Bias-in-Bios` dataset (De-Arteaga et al., 2019). Given the skewed distribution of genders across occupations, models pick up cues that reveal the biographee's gender. For instance, most biographies corresponding to the "professor" occupation are of males. Models trained on this dataset can learn such spurious associations. To measure how much the trained networks encode gender, we train a gender classifier on the internal representations and compute the its minimum description length (MDL) as a measure of the ease of extracting gender from representations (Voita & Titov, 2020).

We consider a variety of training sets by *sub-sampling* and *over-sampling* examples for each profession in the dataset: This is done to balance the number of examples across each gender. We do this for both—the original inputs in the dataset (*raw*) and *scrubbed* examples, wherein gender sensitive information (such as pronouns) is removed (similar to setups in (De-Arteaga et al., 2019)). We perform our analysis on RoBERTa-base (Liu et al., 2019) fine-tuned for these training sets [6].

Results: In Figure 4, we observe the distribution of the two information measures for the last layer of networks trained on the different training sets [7]. This variation is shown in conjunction with *compression* values across the network using the MDL probe. Following our initial hypothesis (Table 1; **H1**), we expect that networks with higher representation of bias will have lower entropy. Indeed, in Figure 4 (left), the network trained on the original training set (i.e., `raw original`) shows the lowest entropy. This finding is in line with our hypothesis, since the other networks are trained on either gender-balanced or scrubbed sets. However, we do not observe consistent trends among networks trained on these de-biasing sets. On the other hand, we do see clear patterns in MI that distinguish networks in line with their compression values (Figure 4, right). As we go from lower to higher values of MI (left to right), the density plots get darker, corresponding to higher compression values (higher bias). A prominent distinction is seen between the `raw` and `scrubbed` sets that are separated on two sides of the plot.

Image Classification with Contexts Next, we consider a scenario from computer vision, where the task is to identify the presented object in a particular context. We use a subset of the dataset from the `NICO`$^{++}$ (Zhang et al., 2022) dataset that consists of images of animals in a variety of contexts.

[6] We use trained checkpoints released by Orgad et al. (2022). More details are given in appendix B.3.

[7] We also make some striking findings on rest of the layers, as discussed in appendix C.2

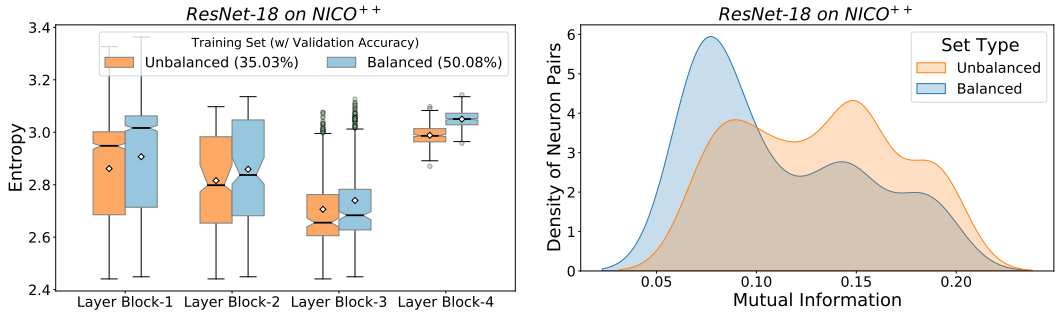

Figure 5: Entropy and MI for ResNet-18 on the NICO^{++} dataset. The two training sets—`balanced` and `unbalanced`—result into models that vary in their generalization to contextual features beyond on what they were trained on. This distinction is reflected in the information measurements.

For each animal class, there exists two types of contexts: *individual*, those that are specific to only that animal and is not present for all classes (such as a *roaring* bear), and *common*, contexts that exist across all classes (such as images taken in *dark*).

For our analysis, we design two training sets—`unbalanced` and `balanced`—varying in the distribution of common contexts across examples. Each animal in the `unbalanced` set occurs in a particular common context that is chosen for that animal. In contrast, the `balanced` set contains images from all common contexts, for each animal. Thus, a network trained on the `unbalanced` set is likely to pick the context-to-animal mapping (i.e., a case of heuristic memorization).

Results: We train ResNet-18 (He et al., 2015) networks for the two sets and evaluate them on the common NICO^{++} evaluation set, balanced across all common contexts. We consider the hidden representation from each of the 4 block of layers in the network to compute the information measures reported in Figure 5. From the left sub-figure, we observe that the entropy for networks trained on the `balanced` set is consistently greater than the `unbalanced` set across all layer blocks. Furthermore, we observe that distribution of MI (right) across pairwise neurons also reflects the difference between the networks, corroborating our hypothesis. Neuron pairs for the network that memorizes the correlation with image contexts (`unbalanced`) are more densely concentrated at higher MI values.

4 Example-level Memorization

We now examine how the distribution of information measures across networks change when they memorize individual examples. Following our original hypotheses (Table 1; **H2**), we expect such networks to display high intra-neuron and inter-neuron diversity, and thus high entropy and low MI.

We perform the analysis for example-level memorization on the standard datasets of `MNIST` (Lecun et al., 1998) and `IMDb` (Maas et al., 2011) on a 3-layered MLP and DistilBERT-base, respectively. In order to study how the diversity in neurons changes with increasing example-level memorization, we induce varying levels of label noise by randomly shuffling a fraction of training examples' target labels (denoted by a parameter β). We then analyze these trained networks on the original validation set.

Results: First, we note that model performance on the validation set decreases with increased label shuffling, validating an increase in example-level memorization (Figure 6). Interestingly, this dip in validation accuracy is accompanied with a consistent rise in entropy across the neurons. For MLP networks trained on `MNIST` (left), we see a distinct rise in entropy even with a small amount of label shuffling ($\beta = 0.25$), followed by a steady increase (layers 2 and 3) or no change (layer 1) in entropy. A dissimilar trend is seen for DistilBERT fine-tuned on `IMDb` (right): a distinct rise for high values of β and a consistent value for low or no label shuffling. While our hypothesis holds true in both settings, we speculate the difference between them is due to the pre-trained initialization of DistilBERT, which has been shown to act as an implicit regularization during fine-tuning (Tu et al., 2020). That is, here,

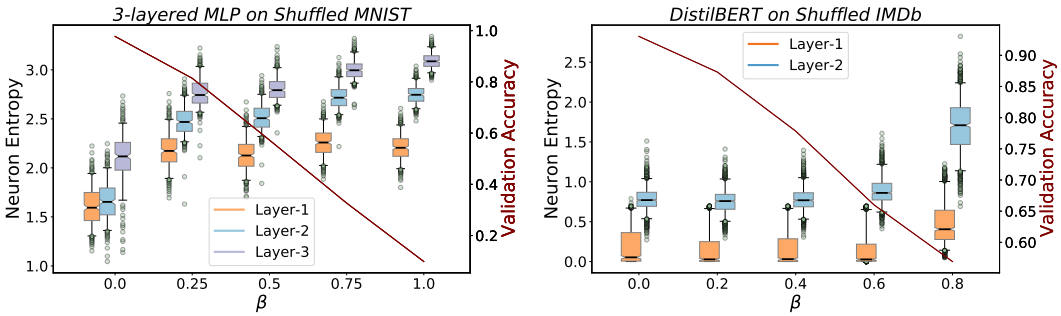

Figure 6: Entropy across neuron activations increases with greater example-level memorization (↑ β).

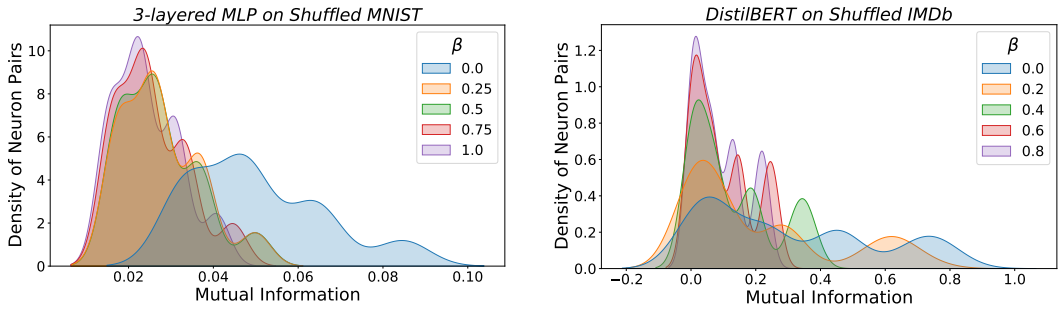

Figure 7: Networks that show higher example-level memorization (↑ β) have high density of neuron pairs for lower MI values. Here, MI is computed across the first layer for both the networks.

DistilBERT might be learning task-relevant information despite some amount of label noise (note that this is not evident through validation performance alone).

Our hypothesis for the relation between example-level memorization and MI is supported by Figure 7. In both settings, networks trained on higher β values consist of neuron pairs that show low values of MI (left side of the plots). In line with the previous observations, we find that for MLPs trained on some amount of label noise (any $\beta > 0.00$) on MNIST, have a higher density of neuron pairs concentrated at low values of MI. Meanwhile, for DistilBERT on `IMDb` (right sub-plot), we observe that neuron pair density gradually shifts towards lower values of MI with increasing β.

5 Model Selection

Table 2: We measure the correlation (Kendall's τ) between model rankings based on their generalization as estimated through extrinsic metrics on labeled test sets and those obtained via information measures. Note that τ can range from -1.0 (perfect disagreement) to 1.0 (perfect agreement).

	Sentiment Adjectives	Colored MNIST	Bias-in-Bios			Shuffled IMDb	Shuffled MNIST
	Validation Accuracy	Validation Accuracy	Comp-ression	TPR Gap	Suff. Gap	Validation Accuracy	Validation Accuracy
Mean Entropy	0.80	1.00	0.47	0.20	0.20	0.60	1.00
Mean MI	0.80	1.00	0.60	0.07	0.33	0.80	0.88

A natural application of our findings is the problem of model selection: given a list of models, rank them based on their generalizability. To demonstrate the utility of our insights, we compare the correlations between rankings obtained through our information-theoretic measures (which do no require labeled data), with the generalization ability of the model on a labeled held-out set.

We consider the same tasks and networks as discussed in the prior sections, and compute the rankings using (i) extrinsic evaluation metrics defined for the task (such as validation accuracy for `Colored MNIST` and compression for `Bias-in-Bios`), (ii) the mean of entropy values, and (iii) the mean of MI values computed for the same networks. We then compute the Kendall rank correlation coefficient, τ, between these rankings (between (i) & (ii), and (i) & (iii)) to evaluate the agreement amongst them.

We observe high correlation values for all the comparisons (Table 2). Particularly high correlations are observed for setups with synthetically induced spurious correlations (§3.1) and shuffled labels (§4), with rankings on `Colored MNIST` being perfectly correlated. Correlations on `Bias-in-Bios` are positive but lower, likely due to the more nuanced setup, where the memorization is less pronounced and extrinsic metrics are weakly correlated among themselves (appendix D.1; Orgad et al., 2022).

6 Related Work

A large body of work aims at measuring and quantifying generalization, especially on out-of-distribution (OOD) scenarios (Ben-David et al., 2010; Hendrycks et al., 2021a; Wang et al., 2021b). The most common approach is to curate and label a set of examples to evaluate if networks exploit certain heuristics or shortcuts (Lapuschkin et al., 2019; Zhao et al., 2018). Several past studies create such sets spanning different domains and tasks to shed light on common failure modes in both the trained models, and the datasets used to train them. A body of such work exists for several tasks across vision (Russakovsky et al., 2015; Hendrycks & Dietterich, 2019; Hendrycks et al., 2021a,b) and NLP (McCoy et al., 2019; Naik et al., 2018; Ravichander et al., 2021; Kim & Linzen, 2020).

Closely related to the motivations for our work, past work has attempted to evaluate models using techniques that go beyond extrinsic evaluation. Training dynamics have been explored to assess the role of individual examples in training sets (Swayamdipta et al., 2020) and how specific knowledge features are temporally picked during training (Saphra & Lopez, 2019). Recent work has noted that frequent spurious artifacts are learnt prior to general patterns during training (Tänzer et al., 2021), in turn followed by memorization of individual examples (Arpit et al., 2017). Closely sharing our motivations of using information-theoretic viewpoints for intrinsic evaluation over network activations, past work has investigated *probing* or *diagnostic* classifiers (Adi et al., 2017; Belinkov et al., 2017; Hupkes et al., 2018). Researchers have further extended this paradigm to identify the role of individual neurons (Dalvi et al., 2019; Durrani et al., 2020; Bau et al., 2017, 2020).

Other work using information theory to study neural networks has focused on their learning process through the lens of the information bottleneck principle (Tishby et al., 1999), categorizing learning into distinct phases (Shwartz-Ziv & Tishby, 2017; Saxe et al., 2018) and obtaining generalization bounds (Tishby & Zaslavsky, 2015). Follow-up work has made use of such measures to regularize training for robustness (Wang et al., 2021a) and low-resource learning (Mahabadi et al., 2021).

7 Conclusion

In this work, we have taken a step towards identifying generalization behavior of neural network models based on their intrinsic activation patterns. We presented information-theoretic measures that allow us to distinguish between models that show two kinds of memorization: those that pick up surface-level spurious correlations (heuristic memorization) and those that overfit on individual training instances (example-level memorization). Through investigations spanning multiple natural language and vision tasks, we corroborated our hypothesis that such memorization is reflected in diversity across neural activations, and hence the defined information measures that quantify them. We also demonstrated a potential application of this framework for model selection.

An exciting avenue for future work is to encourage models to have the expected organization of information during training, for example via regularization. Moreover, these information measures could also be used point-wise to characterize model behavior over individual examples, potentially useful for OOD detection (Arora et al., 2021). There is also room to scale our empirical evaluation further to larger models, as well as other tasks, to study the scaling behavior of these information measures as a function of data and model size (Rosenfeld et al., 2020; Kaplan et al., 2020).

References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJh6Ztuxl.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. URL http://arxiv.org/abs/1907.02893.

Arora, U., Huang, W., and He, H. Types of out-of-distribution texts and how to detect them. In Moens, M., Huang, X., Specia, L., and Yih, S. W. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 10687–10701. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.835. URL https://doi.org/10.18653/v1/2021.emnlp-main.835.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017. URL http://proceedings.mlr.press/v70/arpit17a.html.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, 2017.

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907375117. URL https://www.pnas.org/content/early/2020/08/31/1907375117.

Beaudry, N. J. and Renner, R. An intuitive proof of the data processing inequality. *Quantum Inf. Comput.*, 12(5-6):432–441, 2012. doi: 10.26421/QIC12.5-6-4. URL https://doi.org/10.26421/QIC12.5-6-4.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Hjelm, R. D., and Courville, A. C. Mutual information neural estimation. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 530–539. PMLR, 2018. URL http://proceedings.mlr.press/v80/belghazi18a.html.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. R. What do neural machine translation models learn about morphology? In Barzilay, R. and Kan, M. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 861–872. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1080. URL https://doi.org/10.18653/v1/P17-1080.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. CLUB: A contrastive log-ratio upper bound of mutual information. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1779–1788. PMLR, 2020. URL http://proceedings.mlr.press/v119/cheng20b.html.

Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, A., and Glass, J. R. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 6309–6317. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33016309. URL `https://doi.org/10.1609/aaai.v33i01.33016309`.

Darbellay, G. and Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999. doi: 10.1109/18.761290.

De-Arteaga, M., Romanov, A., Wallach, H. M., Chayes, J. T., Borgs, C., Chouldechova, A., Geyik, S. C., Kenthapadi, K., and Kalai, A. T. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In danah boyd and Morgenstern, J. H. (eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 120–128. ACM, 2019. doi: 10.1145/3287560.3287572. URL `https://doi.org/10.1145/3287560.3287572`.

Durrani, N., Sajjad, H., Dalvi, F., and Belinkov, Y. Analyzing individual neurons in pre-trained language models. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 4865–4880. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.395. URL `https://doi.org/10.18653/v1/2020.emnlp-main.395`.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=Bygh9j09KX`.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=HJz6tiCqYm`.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 8320–8329. IEEE, 2021a. doi: 10.1109/ICCV48922.2021.00823. URL `https://doi.org/10.1109/ICCV48922.2021.00823`.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15262–15271. Computer Vision Foundation / IEEE, 2021b. URL `https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html`.

Hupkes, D., Veldhoen, S., and Zuidema, W. H. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *J. Artif. Intell. Res.*, 61:907–926, 2018.

Kaplan, J., McCandlish, S., Henighan, T. J., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.

Kim, N. and Linzen, T. COGS: A compositional generalization challenge based on semantic interpretation. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 9087–9105. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.731. URL https://doi.org/10.18653/v1/2020.emnlp-main.731.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Phys. Rev. E*, 69: 066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K. Unmasking clever hans predictors and assessing what machines really learn. *CoRR*, abs/1902.10178, 2019. URL http://arxiv.org/abs/1902.10178.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Lee, Y., Lee, J., Hwang, S. J., Yang, E., and Choi, S. Neural complexity measures. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6e17a5fd135fcaf4b49f2860c2474c7c-Abstract.html.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Mahabadi, R. K., Belinkov, Y., and Henderson, J. Variational information bottleneck for effective low-resource fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=kvhzKz-_DMF.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Naik, A., Ravichander, A., Sadeh, N. M., Rosé, C. P., and Neubig, G. Stress test evaluation for natural language inference. In Bender, E. M., Derczynski, L., and Isabelle, P. (eds.), *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pp. 2340–2353. Association for Computational Linguistics, 2018. URL https://aclanthology.org/C18-1198/.

Orgad, H., Goldfarb-Tarrant, S., and Belinkov, Y. How gender debiasing affects internal model representations, and why it matters. *CoRR*, abs/2204.06827, 2022. doi: 10.48550/arXiv.2204.06827. URL https://doi.org/10.48550/arXiv.2204.06827.

Ravichander, A., Dalmia, S., Ryskina, M., Metze, F., Hovy, E., and Black, A. W. NoiseQA: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the*

European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2976–2992, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.259. URL https://aclanthology.org/2021.eacl-main.259.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales. 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=ryenvpEKDr.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Saphra, N. and Lopez, A. Understanding learning dynamics of language models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3257–3267, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1329. URL https://aclanthology.org/N19-1329.

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 *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=ry_WPG-A-.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/2012/papers/625.pdf.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624, 2021. URL https://arxiv.org/abs/2108.13624.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017. URL http://arxiv.org/abs/1703.00810.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 9275–9293. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.746. URL https://doi.org/10.18653/v1/2020.emnlp-main.746.

Tänzer, M., Ruder, S., and Rei, M. BERT memorisation and pitfalls in low-resource scenarios. *CoRR*, abs/2105.00828, 2021. URL https://arxiv.org/abs/2105.00828.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*, pp. 1–5. IEEE, 2015. doi: 10.1109/ITW.2015.7133169. URL https://doi.org/10.1109/ITW.2015.7133169.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

Tu, L., Lalwani, G., Gella, S., and He, H. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 10 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00335. URL https://doi.org/10.1162/tacl_a_00335.

Voita, E. and Titov, I. Information-theoretic probing with minimum description length. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 183–196. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://doi.org/10.18653/v1/2020.emnlp-main.14.

Wang, B., Wang, S., Cheng, Y., Gan, Z., Jia, R., Li, B., and Liu, J. Infobert: Improving robustness of language models from an information theoretic perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=hpH98mK5Puk.

Wang, J., Lan, C., Liu, C., Ouyang, Y., and Qin, T. Generalizing to unseen domains: A survey on domain generalization. In Zhou, Z. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 4627–4635. ijcai.org, 2021b. doi: 10.24963/ijcai.2021/628. URL https://doi.org/10.24963/ijcai.2021/628.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530 [cs]*, February 2017. URL http://arxiv.org/abs/1611.03530. arXiv: 1611.03530.

Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., and Cui, P. Nico++: Towards better benchmarking for domain generalization, 2022. URL https://arxiv.org/abs/2204.08040.

Zhao, Z., Dua, D., and Singh, S. Generating natural adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1BLjgZCb.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] All claims made are supported by experimental evidence. See §3, §4, and §5.
   (b) Did you describe the limitations of your work? [Yes] We discuss limitation throughout the paper while discussing experimental results and in the conclusion (§7).
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See appendix F.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include all resources to reproduce our experimental results in the supplemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See appendix B.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All experiments were run over multiple seeds. All results are shown after averaging over these seeds. Showing error bars was not possible with our results due the nature of visualizations, hence we include relevant details in appendix C.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] Cited throughout the relevant sections in the main paper.

(b) Did you mention the license of the assets? [Yes] We use popular publicly available datasets such as MNIST and IMDb. Yet, we include the licenses for relevant assets in appendix G.

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Valid for the `Bias-in-Bios` dataset; discussed in appendix B.3

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Valid for the `Bias-in-Bios` dataset; discussed in appendix B.3

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]