PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization

Anonymous Author(s)
Affiliation
Address
email

Abstract

While there has been progress in developing non-vacuous generalization bounds for deep neural networks, these bounds tend to be uninformative about why deep learning works. In this paper, we develop a compression approach based on quantizing neural network parameters in a linear subspace, profoundly improving on previous results to provide state-of-the-art generalization bounds on a variety of tasks, including transfer learning. We use these tight bounds to better understand the role of model size, equivariance, and the implicit biases of optimization, for generalization in deep learning. Notably, we find large models can be compressed to a much greater extent than previously known, encapsulating Occam's razor.

1 Introduction

There has been great effort to understand why deep learning models generalize so well on common datasets, despite the fact that these models have many more parameters than the number of training datapoints and can fit random labels [58]. These observations are not explainable through classical statistical learning theory such as VC dimension or Rademacher complexity which focus on uniform convergence over the hypothesis class. The PAC-Bayes framework, in contrast, gives a convenient way of constructing bounds where the generalization gap depends on the deep learning model (or models) that are actually found by training rather than the hypothesis set as a whole. Using this framework, many different potential explanations have been proposed drawing on properties of a deep learning model that are induced by the training dataset, such as low spectral norm [45], noise stability [2], flat minima [28], robustness and compression [2, 59].

In this work, we show that neural networks are substantially more compressible than previously believed when paired with structured training datasets. Constructing tighter generalization bounds than have been previously achieved, we show that this compression *alone* is sufficient to explain much of the generalization of neural networks.

Contributions of this Paper

1. We develop a new approach for training compressed neural networks that adapts the compressed size to the difficulty of the problem. We train in selected random subspaces of the parameters and perform learned quantization. Consequently, we achieve extremely low compressed sizes for neural networks at a given accuracy level, which is a requirement for our tight bounds.
2. Using a prior encoding Occam's razor and our compression scheme, we construct the best generalization bounds to date on image datasets, both with data-dependent and data-independent priors. We also show how transfer learning provides improvements to compression and thus our generalization bounds, explaining the practical performance benefits of pre-training.

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

Table 1: Recent non-vacuous PAC-Bayes bounds obtained on popular image classification datasets in deep learning. $\star$ indicates bounds obtained using data-dependent priors (Section 5.2).

Reference	Non-vacuous PAC-Bayes bounds (%)				
	Binary MNIST	MNIST	FMNIST	CIFAR-10	CIFAR-100
Dziugaite and Roy [17]	16.1	✗	✗	✗	✗
Rivasplata et al. [48]	2.2	✗	✗	✗	✗
Zhou et al. [59]	✗	< 46	✗	✗	✗
Dziugaite et al. [18]	✗	$11^{\star}$	$38^{\star}$	$23^{\star}$	✗
Pérez-Ortiz et al. [47]	✗	$21.7/1.5^{\star}$	✗	$16.7^{\star}$	✗
Our bounds	✗	$\mathbf{11.6/1.4}^{*}$	$\mathbf{32.8/10.1}^{*}$	$\mathbf{58.2/16.6}^{*}$	$\mathbf{94.6/44.4}^{*}$

3. Through the lens of compression and our bounds, we are able to explain why deep learning models generalize on structured datasets like CIFAR-10, but not when structure is broken such as by shuffling the pixels or shuffling the labels, the benefits of equivariant models, and whether the implicit regularization of SGD is necessary for generalization.

2 Related Work

In this section, we delineate recent techniques developed for achieving non-vacuous PAC-Bayes bounds for neural networks.

Optimizing the PAC-Bayes bound Dziugaite and Roy [17] obtain the first non-vacuous generalization bounds for deep stochastic neural networks on binary MNIST. To do so, the authors construct a relaxation of the Langford and Seeger [36] bound and optimize it with the objective of finding a posterior distribution that covers a large volume of low-loss solutions around a local minimum obtained using SGD. Similarly, Rivasplata et al. [48] extend the optimization to other bounds by developing novel relaxations of other PAC-Bayes bounds based on previous work from Blundell et al. [5].

Compressing Models It is well-established that neural networks are robust to small perturbations [27, 28, 35, 34, 33, 45, 7]. The work of Arora et al. [2] leverages this connection to establish a compression-based approach that uses noise stability to study generalization of neural networks. Following this insight, Zhou et al. [59] develops a PAC-Bayes bound that uses the representation of a compressed model in bits, and they also add noise stability through their use of Gaussian posteriors and Gaussian mixture priors. Furthermore, they achieve the smallest model representations through pruning and quantization following ideas of Han et al. [26] and Cheng et al. [9].

Data-Dependent Priors Dziugaite et al. [18] demonstrated that for linear PAC-Bayes bounds such as Thiemann et al. [51], a tighter bound can be achieved by choosing the prior distribution to be data-dependent. That is, the prior is trained to concentrate around low loss regions on held-out data. More precisely, the authors show that the optimal data-dependent prior is the conditional expectation of the posterior given a subset of the training data. They approximate this data-dependent prior by solving a variational problem over Gaussian distributions. They evaluate the bounds for SGD-trained networks on data-dependent priors obtaining tight bounds on MNIST, Fashion MNIST, and CIFAR-10. In a similar vein, Pérez-Ortiz et al. [47] combine the data-dependent priors approach Dziugaite et al. [18] with the PAC-Bayes with Backprop (PBB) approach from Rivasplata et al. [48] to obtain *state-of-the-art* PAC-Bayes non-vacuous bounds for MNIST and CIFAR-10 with data-dependent priors. Table 1 summarizes the PAC-Bayes bounds obtained by these works alongside our bounds.

Our upper bound on the KL term closely follows Zhou et al. [59], but with key improvements: (i) we train in lower dimensional subspaces using intrinsic dimensionality and FiLM subspaces which proves to be more effective and adaptable than pruning, (ii) we consider a more effective quantization scheme with variable length codes and quantization aware training, and (iii) we take advantage of the increased compression provided by transfer learning and data dependent priors. As a result, we obtain state-of-the-art non-vacuous bounds for MNIST, Fashion MNIST, CIFAR-10, and CIFAR-100 in both the from scratch and transfer learning setting as shown in Section 4. Motivated by the tightness of these bounds, we seek to better understand the role of equivariance, stochastic optimization, and data dependent bounds, for explaining generalization.

Ding et al. [15] is another closely related work which optimizes the PAC-Bayes bounds to predict the transferability of various upstream models to a downstream task. In contrast, we aim to explain why neural networks generalize through compression. This is different than predicting transferability as there are several factors other than compression (which we demonstrate is sufficient, but not necessary for generalization) that could make a model perform well on a downstream task.

3 PAC-Bayes Bounds

PAC-Bayes bounds are fundamentally an expression of Occam's razor: simpler descriptions of the data generalize better. As an illustration, consider the finite hypothesis generalization bound. Let $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(h\left(x_i\right), y_i\right)$ be the empirical risk of a hypothesis h from a finite hypothesis set $\mathcal{H}$ for the 0-1 loss ℓ, and $R(h) = \mathbb{E}[\hat{R}(h)]$ its population risk. With probability at least $1-\delta$, the population risk for a hypothesis h using n data samples satisfies

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\log|\mathcal{H}| + \log(1/\delta)}{2n}}, \tag{1}$$

bounded by the empirical risk and a complexity term $\log|\mathcal{H}|$ counting the number of bits needed to specify any $h \in \mathcal{H}$.

But what if we don't believe that each hypothesis is equally likely? If we consider that a given prior distribution P over $\mathcal{H}$ concentrates around the optimal hypothesis, then we can construct a variable length code that uses fewer bits for the more likely hypothesis. With the best possible code for P, a given hypothesis h will take $\log_2 \frac{1}{P(h)}$ bits to represent. Using this prior may allow us to pay for a smaller complexity term even if the hypothesis set is large, so long as the hypotheses that are consistent with the data are also likely under the prior.

The number of bits payed can further be reduced from $\log_2 \frac{1}{P(h)}$ to $\mathbb{KL}(Q \parallel P)$ by considering a distribution of "good" solutions Q. If we don't care which element of Q we arrive at, we can gain some bits *back* from this insensitivity (which could be used to code a separate message). The average number of bits to code a sample from Q using the prior P is $\mathbb{H}(Q, P)$, and one gets $H(Q)$ bits back from being agnostic about which sample $h \sim Q$ to use, yielding the KL-divergence between Q and P: $\mathbb{H}(Q, P) - \mathbb{H}(Q) = \mathbb{KL}(Q \parallel P)$.

With these improvements of the finite hypothesis bound: replacing $\log|\mathcal{H}|$ with $\mathbb{KL}(Q \parallel P)$, and sampling a hypothesis $h \in \mathcal{H}$, one arrives (with minor bookkeeping) at the PAC-Bayes bound introduced in McAllester [43]. With probability at least $1-\delta$,

$$\mathop{\mathbb{E}}_{h\sim Q}\left[R\left(h\right)\right] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\hat{R}\left(h\right)\right] + \sqrt{\frac{\mathbb{KL}(Q \parallel P) + \log(n/\delta) + 2}{2n-1}}. \tag{2}$$

Many refinements of Eq. (2) have been developed [36, 42, 6, 51] but retain the same character. The lower the ratio of the $\mathbb{KL}$ to the number of data points n, the tighter the gap between empirical and expected risk. In this work, we use the tighter Catoni [6] variant of the PAC-Bayes bound (see Appendix I for details).

Universal Prior. Connecting back to Occam's razor, one can define a prior that explicitly penalizes the minimum compressed length of the hypothesis, also known as the universal prior [50]: $P(h) = 2^{-K(h)}/Z$, where K is the *prefix* Kolmogorov complexity [30] of h (the length of the shortest program that produces h and also delimits itself), and $Z \leq 1$.[1] Using a point mass posterior on a single hypothesis h^*, $\mathbb{KL}\left(Q \parallel P\right)$ from the bound reduces to

$$\mathbb{KL}\left(\mathbf{1}_{[h=h^*]} \parallel P(h)\right) = \log\tfrac{1}{P(h)} \leq K(h)\log 2 \leq l(h)\log 2 + 2\log l(h),$$

where $l(h)$ is the length of a given program that reproduces h not including the delimiter, which we can use when upper bounding the KL. For convenience, we can condition on using the same method for compression and decompression for all elements of the prior. If we can reduce the compressed

[1] The universal prior similar to the discrete hypothesis prior from Zhou et al. [59] but setting $m(h) = 2^{-2\log_2 l(h)}$ rather than the flat $m(h) = 2^{-72}$.

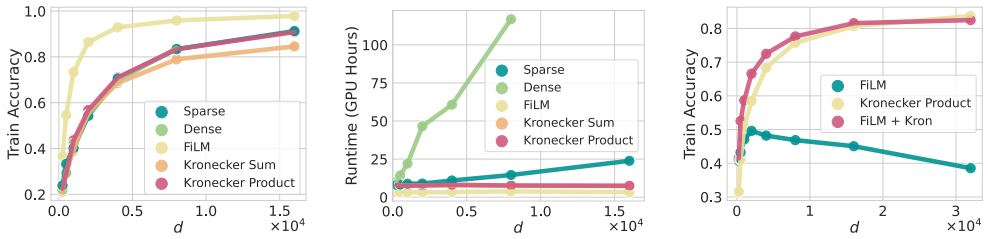

Figure 1: **(Left)** Different projection operators P (Section 4.1) used for transfer learning from Imagenet to CIFAR-10 on a ResNet-34 across different subspace dimensions d. Kronecker product, Sparse, and Dense perform almost identically **(Center)** Kronecker product runs with substantially reduced the runtime cost compared to the Sparse or Fastfood matrices used by Li et al. [38]. **(Right)** Training from scratch on CIFAR-10. The FiLM projector alone is unable to fit the data when training from scratch, and instead a sum of FiLM and Kronecker Product projectors perform the best.

length $l(h)$ of the hypotheses we find that fits the training data, then we can improve the tightness of the PAC-Bayes bound.

Model Compression. *Model compression* aims to find a nearly equivalent model that can be expressed fewer bits and is useful for deployments on mobile devices or for improving inference time on specialized hardware [8, 9]. For computing PAC-Bayes generalization bounds, however, only the model size matters. Therefore, we can employ compression methods which may otherwise be unfavorable in practice due to worse computational requirements. *Pruning* and *quantization* are among the most widely used methods for model compression. In this work, we rely on quantization (Section 4.2) to achieving tighter generalization bounds.

4 Tighter Generalization Bounds via Adaptive Subspace Compression

Training a neural network involves taking many gradient steps in a high-dimensional space $\mathbb{R}^D$. Although D may be large, the loss landscape has been, in some sense, found to be simpler than typically believed [14, 25, 22]. Analogous to the notion of intrinsic dimensionality more generally, Li et al. [38] attempt to find the lowest dimensional subspace in which the network can be trained and still fit the training data. The weights of a neural network $\theta \in \mathbb{R}^D$ are parametrized in terms of an initialization θ_0 and a projection $w \in \mathbb{R}^d$ to a lower dimensional subspace through a fixed matrix $P \in \mathbb{R}^{D \times d}$,

$$\theta = \theta_0 + Pw. \tag{3}$$

To facilitate favorable conditioning during optimization, P is chosen to be approximately orthonormal $P^\top P \approx I_{d \times d}$ and for scalability, Li et al. [38] utilize random normal matrices of the form $P \sim \mathcal{N}(0,1)^{D \times d}/\sqrt{D}$, and their sparse approximations [39, 37].

In its original form, intrinsic dimensionality (ID) is used as a scientific tool to measure the complexity of the learning task. Unlike other methods like pruning, ID has the benefit of scaling dramatically up or down in size with difficulty of the problem. While ID was not used with quantization as a full fledged model compression method, we find that it can be. For our work, the ability to find the intrinsic dimension $d \ll D$ has profound implications for the compressibility of the models, and therefore our ability to construct generalization bounds. As demonstrated by Zhou et al. [59], the compressibility of neural network models has a direct connection to generalization and immediately allows us to compute non-vacuous PAC-Bayes bounds for transfer learning. Building upon ID, we describe our approach next.

4.1 Finding Better Random Subspaces to Train In

To further improve upon the scalability and effectiveness of the projections P used by Li et al. [38], we introduce three new methods.

Kronecker Sum Projector: Using the Kronecker product $\otimes$, we construct the matrix $P_\oplus = (\mathbf{1} \otimes R_1 + R_2 \otimes \mathbf{1})/\sqrt{2D}$ where $R_1, R_2 \sim \mathcal{N}(0,1)^{\sqrt{D} \times d}$ and $\mathbf{1}$ is the vector of all ones

in $\mathbb{R}^{\sqrt{D}}$. From the fact that $R_1 \perp\!\!\!\perp R_2$ and the entries are standard normal, one can verify that $P_{\oplus}^{\top} P_{\oplus} = I_{d\times d} + \mathcal{O}(1/\sqrt{D})$.

Kronecker Product Projector: Alternatively, we can form the matrix $P_{\otimes} = Q_1 \otimes Q_2/\sqrt{D}$ with the smaller $Q_1, Q_2 \sim \mathcal{N}(0,1)^{\sqrt{D}\times\sqrt{d}}$, and again this matrix is approximately orthogonal: $P_{\otimes}^{\top} P_{\otimes} = (Q_1^{\top} Q_1/\sqrt{D}) \otimes (Q_2^{\top} Q_2/\sqrt{D}) = I \otimes I + O(D^{-1/4}) = I_{d\times d} + O(D^{-1/4})$. [2]

The matrix vector multiply $w \mapsto Pw$ for both the above projectors can be performed in time $O(d\sqrt{D})$ and $O(\sqrt{dD})$ respectively, rather than the $O(dD)$ that is required by the dense random matrix. Figure 1 demonstrates the runtime speedup and training performance improvement in comparison to the methods using by Li et al. [38]. Notably, the Kronecker-structured projections retain the fidelity of the dense random matrix while being orders of magnitude faster than the alternative operators when scaling to larger values of d.

FiLM projector: It is well known that the BatchNorm parameters have an outsized effect on the downstream task performance relative to their size, and this fact has been used in Featurewise independent Linear Modulation (FiLM) [46, 16] for efficient control of neural networks in many different settings. Several authors have explored performing finetuning for transfer learning solely on these parameters and the final linear layer [32]. Drawing on these observations, we construct a projection matrix P_{FiLM} where only columns corresponding with BatchNorm or head parameters are non-zero and sampled from $\mathcal{N}(0,1)^d/\sqrt{D}$, which we also show in Figure 1. While the FiLM projector is highly effective for transfer learning (shown in Figure 1 left), the performance saturates quickly when training from scratch. For this reason, when training from scratch we employ the sum $P_{\text{FiLM}+\otimes} = (P_{\text{FiLM}} + P_{\otimes})/\sqrt{2}$, which outperforms the two projectors individually as shown in Figure 1 right.

4.2 Quantization Scheme and Training

As a consequence of finite-precision computations (e.g. 32-bit floating points), the hypothesis space $\mathcal{H}$ of any neural network remains finite in practice. Moreover, the ability of neural networks to be compressed [26, 11] to far fewer number of bits per parameter while retaining similar performance allows us to compress the learned subspace vector $w \in \mathbb{R}^d$ from Eq. (3) as well.

Given the full precision weights $w = [w_1, \ldots, w_d] \in \mathbb{R}^d$ and a vector $c = [c_1, \ldots c_L] \in \mathbb{R}^L$ of L quantization levels, we construct the quantized vector $\hat{w} = [\hat{w}_1, \ldots, \hat{w}_d]$ such that $\hat{w}_i = c_{q(i)}$ where $q(i) = \arg\min_k |w_i - c_k|$. The quantization levels c are learned alongside w, where the gradients are defined using the straight through estimator [3, 56]:

$$\frac{\partial \hat{w}_i}{\partial w_j} = \delta_{ij} \quad \text{and} \quad \frac{\partial \hat{w}_i}{\partial c_k} = 1_{[q(i)=k]} \tag{4}$$

We initialize c with uniform spacing between the minimum and maximum values in parameter vector w, which performs better than k-means initialization, as also observed in Choi et al. [10]. To further compress the network, we use a variable length code in the form of arithmetic coding [41], which takes advantage of the fact that certain quantization levels are more likely than others. Given probabilities p_k (empirical fractions) for cluster c_k, arithmetic coding of w takes at most $\lceil d \times \mathbb{H}(p) \rceil + 2$ bits, where $\mathbb{H}(p)$ is the entropy $\mathbb{H}(p) = -\sum_k p_k \log_2 p_k$. For a small number of quantization levels, we find it makes a small improvement in the compression over more common Huffman coding.

In summary, we use $\lceil d \times \mathbb{H}(p) \rceil + 2$ bits for coding the quantized weights $\hat{w}$, $16L$ bits for the codebook c (represented in half precision), and additional $L \times \lceil \log_2 d \rceil$ bits for representing the probabilities p_k, arriving at $l(w) \leq \lceil d \times \mathbb{H}(p) \rceil + L \times (16 + \lceil \log_2 d \rceil) + 2$.

[2] As neither D nor d is typically a perfect square, we concatenate a dense random matrix to pad out the difference between D, d, and a perfect square. As $(\sqrt{D}+1)^2 = D + 2\sqrt{D} + 1$, we have that the size of this padding is at most $\sqrt{D} \times \sqrt{d}$, so it does not increase the asymptotic cost of performing the matrix vector multiplies.

Algorithm 1 Compute PAC-Bayes Bound.

Inputs: Neural network f_θ, Training dataset $\{x_i, y_i\}_{i=1}^n$, Clusters L, Intrinsic dimension d, Confidence $1-\delta$, and Prior distribution P.
function COMPUTE_BOUND($f_\theta, L, d, (x_i, y_i)_{i=1}^n, \delta, P$)
  $w \leftarrow$ TRAIN_ID(f_θ, d, $(x_i, y_i)_{i=1}^n$) ▷ (Section 4.1)
  $\hat{w} \leftarrow$ TRAIN_QUANTIZE(w, L, $(x_i, y_i)_{i=1}^n$)
  Compute quantized train error $\hat{R}(\hat{w})$.
  $\mathbb{KL}(Q, P) \leftarrow$ GET_KL($\hat{w}, P$) ▷ (Section 3)
  return GET_CATONI_BOUND($\hat{R}(\hat{w})$, $\mathbb{KL}(Q, P)$, δ, n) ▷ (Section 3)
end function
function TRAIN_QUANTIZE($w, L, (x_i, y_i)_{i=1}^n$) ▷ (Section 4.2)
  Initialize $c \leftarrow$ GET_CLUSTERS(w, L)
  for $i = 1$ to quant_epochs **do**
    $c \leftarrow c - \rho \nabla_c \mathcal{L}(w, c)$ and $w \leftarrow w - \rho \nabla_w \mathcal{L}(w, c)$
  end for
  return $\hat{w}$
end function
function GET_KL($(\hat{w}, P)$)
  c, count $\leftarrow$ GET_UNIQUE_VALS_COUNTS($\hat{w}$)
  message_size $\leftarrow$ DO_ARITHMETIC_ENCODING($\hat{w}$, c, count)
  message_size $\leftarrow$ message_size + hyperparam_search ▷ (Section 4.3)
  return message_size $+ 2 \times \log$ (message_size)
end function =0

4.3 Subspace Dimension Optimization and Hyperparameters in the Universal Prior

The smaller the chosen intrinsic dimension d, the more similar θ is to the initialization θ_0 in Eq. (3). Consequently, that value of θ is more likely under the universal prior given the shorter description length. Note that in this prior, we condition on the random seed used to generate θ_0 and P. As we optimize over different parameters such as the subspace dimension $d = 1, .., D$, and possibly other hyperparameters such as the learning rate, or number of quantization levels L, we must encode these into our prior and thus pay a penalty for optimizing over them. We can accomodate this very simply by considering the hypothesis h as not just specifying the weights, but also specifying these hyperparameters: $h = (\theta, d, L, \mathrm{lr})$, and therefore using the universal prior $P(h) = 2^{-K(h)}/Z$ we pay additional bits for each of these quantities: $K(h) \leq K(\theta|d, L) + K(d) + K(L) + K(\mathrm{lr})$. If we optimize over a fixed number H of distinct values known in advance for a given hyperparameter such as L, then we can code L using this information in only $\log_2(H)$ bits. In general, we can also bound the dimensionalities searched over by the maximum D so that $K(d) \leq \lceil \log_2 D \rceil$ in any case.

For transfer learning, we replace θ_0 with a learned initialization $\theta_{\mathcal{D}}$ that is found using the pretraining task and data $\mathcal{D}$. With the ID compression, the universal prior $P(h|\mathcal{D})2^{-K(h|\mathcal{D})}/Z$ will place higher likelihood on solutions θ that are close to the pretraining solution $\theta_{\mathcal{D}}$.

5 Empirical Non-Vacuous Bounds

5.1 Non-Vacuous PAC-Bayes Bounds

To compute our bounds on benchmark datasets, we use the joint approach of projected subspace training, parameter quantization, and optimizing the subspace dimension with respect to the bounds as summarized in Algorithm 1. We also optimize other bound hyperparameters such as the learning rate of the quantization-aware training. We offer more details about bound optimization in Appendix D. Our results for *data-independent* priors, i.e., priors that do are not pre-trained on a subset of the training data, are shown in Table 2. We derive the first non-vacuous bounds on FashionMNIST, CIFAR-10, and CIFAR-100 without data-dependent priors. These results are of significance as we argue in Section 5.2 that using data-dependent priors are not explanatory about the learning process. In particular, we improve over the compression bound results obtained by Zhou et al. [59] on MNIST from $< 46\%$ to 11.55%. We also dramatically improve their compression results as we reduce the compressed size for the best bound by 94% bringing it down from 6.23 KB to 0.38KB for MNIST with LeNet5. The tightness of our SOTA subspace compression bounds allows

Table 2: Our PAC-Bayesian Subspace Compression Bounds with data-independent priors compared to *state-of-the-art* (SoTA) bounds. All results are obtained with 95% confidence, i.e. $\delta = .05$.

Dataset	Data-independent priors		Data-dependent priors	
	Err. Bound (%)	SoTA (%)	Err. Bound (%)	SoTA (%)
MNIST	**11.6**	21.7 [47]	**1.4**	1.5 [47]
+ SVHN Transfer	**9.0**	–	–	–
FashionMNIST	**32.8**	–	**10.1**	38 [18]
+ CIFAR-10 Transfer	**28.2**	–	–	–
CIFAR-10	**58.2**	–	**16.6**	16.7 [47]
+ ImageNet Transfer	**35.1**	–	–	–
CIFAR-100	**94.6**	–	**44.4**	–
+ ImageNet Transfer	**81.3**	–	–	–

us to improve the understanding of several deep learning phenomena as discussed in Section 6. See Appendix A for additional results.

5.2 Data-Dependent Priors

So far, we demonstrated the strength of our bounds on *data-independent* priors, where we considerably improve on the state-of-the-art. However, a number of recent papers have considered data-dependent priors as a way of achieving tighter bounds [47, 18]. In this setup, the training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is partitioned into two parts, $\mathcal{D}_a$ and $\mathcal{D}_b$, with length $n - m$ and m. The first dataset is used to construct a data dependent prior $P(h|\mathcal{D}_a)$, and then the bound is formed over the remaining part of the process: the adaptation of the prior $P(h|\mathcal{D}_a)$ to the posterior $Q(h)$ using the data $\mathcal{D}_b$. The empirical risk is computed over $\mathcal{D}_b$ only. Intuitively, using dataset $\mathcal{D}_a$ it is possible to construct a much tighter prior over the possible neural network solutions.

In our setting, similar to transfer learning, we use the prior $P_{\mathcal{D}_a}(\theta) = 2^{-K(\theta|\theta_{\mathcal{D}_a})}/Z$ where for compression we use $\theta = \theta_{\mathcal{D}_a} + Pw$, and $\theta_{\mathcal{D}_a}$ is the solution found by training the model (without random projections) on the data $\mathcal{D}_a$ rather than initializing randomly. With these data-dependent priors, we achieve the best bounds in Table 2.

However, our adaptive approach exposes a significant downside of data-dependent priors. To the extent that PAC-Bayes bounds can be used for explanation, in this case they only yield insights into the procedure used to adapt $P_{\mathcal{D}_a}(\theta)$ to Q using $\mathcal{D}_b$, any learning that is done in finding $P_{\mathcal{D}_a}(\theta)$ is not constrained or explained by the bound (we have no information on the compressibility of $\theta_{\mathcal{D}_a}$). Given the ability to adapt the size of the KL to the difficulty of the problem, it is possible to squeeze all of the learning into $P_{\mathcal{D}_a}(\theta)$ and none in this adaption to Q. This happens as the $KL \to 0$, which we find happens empirically (or very nearly so) for many splits of the data where $n - m$ is large. Setting $Q(\theta) = \mathbf{1}_{[\theta=\theta_{\mathcal{D}_a}]}$, the KL has only the contribution from the optimization over d: $\mathbb{KL}(Q||P_{\mathcal{D}_a}) \leq \log D$ we find that the bound is nothing more than a variant of the simple Hoeffding bound where $\mathcal{D}_b$ is the validation set $R\left(\theta_{\mathcal{D}_a}\right) \leq \hat{R}_{\mathcal{D}_b}\left(\theta_{\mathcal{D}_a}\right) + \sqrt{\frac{\log(Dm/\delta)+2}{2m-1}}$.

This phenomenon can be seen in Appendix C, where optimization over $d = 1, ..., D$ leads to very small values of the KL when $n - m$ is a large fraction of n, and hence highly simple solutions approaching the Hoeffding bound or confidence interval. This demonstrates that data-dependent priors based PAC-Bayes bounds are not particularly meaningful, and attempts to optimize these bounds will lead inevitably to the statistical bounds formed from the validation set such as a simple Hoeffding bound or confidence interval. While these bounds are correct, they are only informative about the difference in performance of the pretrained model on validation data and test data.

Note that this is in contrast with the data-independent bounds that constrain and explain generalization for the entirety of the learning process. Similarly, our transfer learning bounds meaningfully constrain what happens in the finetuning on the downstream task, but they do not constrain the prior determined from the upstream task.

5.3 Non-Vacuous PAC-Bayes Bounds for Transfer Learning

By directly interpreting PAC-Bayes bounds through the lens of compression, we immediately see the benefits of employing an upstream dataset for transfer learning. In fact, transfer learning allows us to constrain the prior around parameters we know are consistent with the upstream task, and hopefully the downstream task as well, reducing the KL-divergence between the prior and the posterior and leading to even tighter bounds as we show in Table 2. We obtain unequivocally tighter bounds using transfer learning for data-independent priors, providing a theoretical certification that transfer learning can improve generalization. Our PAC-Bayes transfer learning approach also indicates that transfer learning can boost generalization whenever codings optimized on a pre-training task are more efficient for encoding a downstream posterior than an a priori guess made before seeing data. In contrast, downstream tasks which differ greatly from the upstream task may only be consistent with models that are not compressible under the learned prior, a scenario which may describe negative transfer. As pre-training becomes the default in deep learning, our bounds open the door to providing learning guarantees for state-of-the-art techniques. See Appendix B for more details.

6 Understanding Generalization through PAC-Bayes Bounds

Unlike the classical statistical learning theory viewpoint of uniform convergence, where the focus is on the properties of the hypothesis class as a whole, the ability of a given network to be compressed is highly dependent on the actual dataset that the model was trained on. In this respect, the PAC-Bayes perspective shows us that the ability to compress and therefore generalize is not a function of the hypothesis space but rather of individual functions: there are many networks in that space that fit corrupted data and provably cannot be compressed (see Appendix A.1). But these models are not the ones that we find when training on real world datasets, as real world datasets have a tremendous amount of structure. If they did not, it would be pointless trying to learn from them as famously argued by Hume [29] and the No Free Lunch theorem [54, 24]. When training on these structured datasets that have low Kolmogorov complexity, the compressibility of the dataset is reflected in the compressiblity of the model. Breaking the structure in the dataset also breaks the structure in the model and leads to worse generalization bounds for shuffled labels, shuffled pixels, and equivariance more generally.

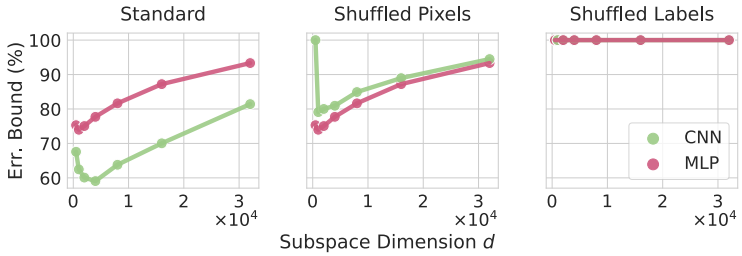

Figure 2: Our PAC-Bayes Bound vs subspace dimension for a fixed size convolutional network and MLP both with 500K parameters. With training on **(Left)** Standard CIFAR-10 **(Middle)** CIFAR-10 with shuffled pixels **(Right)** CIFAR-10 with shuffled labels. Structure in the dataset induces structure in the model. As structure is removed from the dataset, the models become much less compressible hence generalize worse. When the image structure is destroyed with pixel shuffling, we find that the CNN generalizes worse than an MLP.

MLPs vs CNNs: It is well known that convolutional neural networks (CNNs) generalize much better than standard multilayer perceptrons (MLPs) even when controlling for the number of parameters. We find that this fact is reflected in the improved compressibility of CNNs and our generalization bounds, shown in Figure 2 (a) on the image dataset CIFAR-10 using the same number of parameters for the two models. The CNN is able find a more structured and lower description length explanation of the training data.

Shuffled Pixels: However, when this image structure is broken by shuffling the pixels we find that CNNs are no better at generalizing than MLPs, and this is reflected in the compressed sizes. The CNNs become substantially less compressible when trained on this dataset, and hence our bounds show them generalizing worse than MLPs, see Figure 2 (b). Not using this image structure (and

invariant to a global permutation of the features), MLPs do not suffer when this structure is broken since they never used it in the first place.

Shuffled Labels: When the structure of the dataset is entirely broken by shuffling the labels, the compressibility of the models (both for CNNs and MLPs) are lost. Regardless of the subspace dimension used, our generalization bounds are all at 100% error. Using low subspace dimensions it is not possible to fit the training data, and when using a large enough dimension to fit the data the compressed size of the model is larger than the training data and hence the generalization bounds are vacuous.

Equivariance: Designing models which are *equivariant* to certain symmetry transformations has been a guiding principle for the development of data efficient neural networks in many domains [12, 13, 52, 53, 21, 31]. While intuitively it is clear that respecting dataset symmetries severely improves generalization, relatively little has been proven for neural networks [20, 60, 4, 19]. We compress and evaluate rotationally equivariant and non-equivariant Wide ResNets [53, 57] trained on MNIST and a rotated version of MNIST. As shown in Appendix E, the rotationally equivariant models are more compressible and provably generalize better than their non equivariant counterparts when paired with a dataset that also has the rotational symmetry.

Is Stochasticity Necessary for Generalization? It is widely hypothesized that the implicit biases of SGD help find solutions which generalize better. For example, Arora et al. [1] argue that there is no regularizing function with a gradient that replicates the benefits of gradient noise. Wu et al. [55], Smith et al. [49], and Li et al. [40] advocate that gradient noise is necessary to achieve state-of-the-art performance. In contrast, recent work by Geiping et al. [23] shows that full-batch gradient descent can match state-of-the-art performance.

We train ResNet-18 and LeNet5 models on CIFAR-10 and MNIST, respectively, using full-batch and stochastic gradient descent with different values of the intrinsic dimensionality. We provide the training details in Appendix F, but it is worth noting that, unlike Geiping et al. [23], we do not use flatness inducing regularization for full batch training. For MNIST with LeNet5, the best generalization bounds that we obtain are 11.55% and 11.20% using stochastic gradient descent (SGD) and full batch training respectively. The best generalization bounds that we obtain for CIFAR-10 with ResNet-18 are 74.68% and 76.25% using SGD and full batch training, respectively. These close theoretical guarantees on the generalization error for both SGD and full batch training suggest that the implicit biases of SGD may be helpful, but are not at all crucial to understanding why neural networks generalize well. We expand this analysis in Appendix F.

Double Descent [44]: In Appendix H, we show that our bounds are tight and fine-grained enough to predict the double descent phenomenon, and we discuss the significance of this result.

7 Discussion

In this work, we have constructed a new method for compressing deep learning models which is highly adaptive to the level of structure in the model and the training dataset. According to the Occam's prior which considers shorter compressed length models to be more likely, we construct state-of-the-art generalization bounds across a variety of settings. Through our bounds and the viewpoint of Occam's razor and compressibility, we show how generalization relates to the structure in the dataset and the structure in the model, and we are able to explain aspects of neural network generalization for natural image datasets, shuffled pixels, shuffled labels, equivariant models, and for stochastic training.

Limitations We note that despite the power of this compression scheme and our bounds, there are certain aspects of deep learning generalization that remain unexplained. Our compression bounds still show a preference towards models with a smaller number of base parameters as shown in Appendix G, despite the fact that larger models tend to generalize better. While we have achieved better model compression than previous works, it is unlikely that we are close to theoretical limits. We hope that with a further improved compression scheme (possibly using nonlinear parameter projections), we might find that larger deep learning model solutions are compressible to smaller sizes than smaller models. Additionally, while our bounds show that the compressibility of our models implies generalization, we have no statements in the reverse direction. We believe however that the compression approach and Occam's razor have yet untapped explanatory power in deep learning.

References

[1] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. 9

[2] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018. 1, 2

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or Propagating Gradients through Stochastic Neurons for Conditional Computation. *Preprint arXiv:1308.3432v1*, 2013. 5

[4] Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. *Advances in Neural Information Processing Systems*, 34, 2021. 9

[5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 2

[6] Olivier Catoni. PAC-Bayesian Supervised Classification: the Thermodynamics of Statistical Learning. *Institute of Mathematical Statistics Lecture Notes*, 2007. 3, 15

[7] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019 (12):124018, 2019. 2

[8] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017. 4

[9] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136, 2018. 2, 4

[10] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Towards the limit of network quantization. *arXiv preprint arXiv:1612.01543*, 2016. 5

[11] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Towards the Limit of Network Quantization. *The 6th International Conference on Learning Representations (ICLR)*, 2017. 5

[12] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 9

[13] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018. 9

[14] Yann Dauphin, Razvan Pascanu, Çaglar Gülçehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS*, 2014. 4

[15] Nan Ding, Xi Chen, Tomer Levinboim, Beer Changpinyo, and Radu Soricut. Pactran: Pac-bayesian metrics for estimating the transferability of pretrained models to classification tasks. *arXiv preprint arXiv:2203.05126*, 2022. 3

[16] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 3(7):e11, 2018. 5

[17] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017. 2

[18] Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Aprino, and Daniel M. Roy. On the Role of Data in Pac-Bayes Bounds. *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021. 2, 7

[19] Bryn Elesedy. Group symmetry in pac learning. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022. 9

[20] Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International Conference on Machine Learning*, pages 2959–2969. PMLR, 2021. 9

[21] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pages 3165–3176. PMLR, 2020. 9

[22] T. Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *NeurIPS*, 2018. 4

[23] Jonas Geiping, Micah Goldblum, Phillip E. Pope, Michael Moeller, and Tom Goldstein. Stochastic Training Is Not Necessary For Generalization. *The 10th International Conference on Learning Representations (ICLR)*, 2022. 9

[24] Christophe Giraud-Carrier and Foster Provost. Toward a justification of meta-learning: Is the no free lunch theorem a show-stopper. In *Proceedings of the ICML-2005 Workshop on Meta-learning*, pages 12–19, 2005. 8

[25] Ian J. Goodfellow and Oriol Vinyals. Qualitatively characterizing neural network optimization problems. *CoRR*, abs/1412.6544, 2015. 4

[26] Song Han, Huizi Mao, and William J. Dally. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *The 4th International Conference on Learning Representations (ICLR)*, 2016. 2, 5

[27] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993. 2

[28] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. 1, 2

[29] David Hume. *A Treatise of Human Nature*. Oxford University Press, Oxford, 1978. revised P.H. Nidditch. 8

[30] Marcus Hutter et al. Algorithmic complexity. 2008. 3

[31] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 9

[32] Fahdi Kanavati and Masayuki Tsuneki. Partial transfusion: on the expressive influence of trainable batch norm parameters for transfer learning. In *Medical Imaging with Deep Learning*, pages 338–353. PMLR, 2021. 5

[33] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. 2

[34] John Langford. *Quantitatively tight sample complexity bounds*. PhD thesis, Carnegie Mellon University, 2002. 2

[35] John Langford and Rich Caruana. (not) bounding the true error. *Advances in Neural Information Processing Systems*, 14, 2001. 2

[36] John Langford and Matthias Seeger. Bounds for Averaging Classifiers. *Tech. rep CMU-CS-01-102, Carnegie Mellon University*, 2001. 2, 3, 15

[37] Quoc V. Le, Tamás Sarlós, and Alex Smola. Fastfood: Approximate kernel expansions in loglinear time. *ArXiv*, abs/1408.3060, 2013. 4

[38] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *ArXiv*, abs/1804.08838, 2018. 4, 5

[39] Ping Li, Trevor J. Hastie, and Kenneth Ward Church. Very sparse random projections. In *KDD '06*, 2006. 4

[40] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the Validity of Modeling SGD with Stochastic Differential Equations (SDEs). *Preprint arXiv:2102.12470*, 2021. 9

[41] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 5

[42] Andreas Maurer. A Note on the PAC Bayesian Theorem. *Preprint arXiv 041099v1*, 2004. 3, 15

[43] David A McAllester. PAC-Bayesian Model Averaging. *Proceedings of the 12th Annual Conference on Learning Theory (COLT)*, pages 164–170, 1999. 3, 15

[44] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021, 2020. 9

[45] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017. 1, 2

[46] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 5

[47] María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepersvári. Tighter Risk Certificates for Neural Networks. *Journal of Machine Learning Research 22 (2021) 1-40*, 2021. 2, 7

[48] Omar Rivasplata, Vikram M Tankasali, and Csaba Szepesvári. Pac-bayes with backprop. *arXiv preprint arXiv:1908.07380*, 2019. 2

[49] Samuel L. Smith, Erich Elsen, and Soham De. On the Generalization Benefit of Noise in Stochastic Gradient Descent. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. 9

[50] Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7 (1):1–22, 1964. 3

[51] Niklas Thiemann, Christian Igel, Oliver Wintenberger, and Yevgeny Seldin. A Strongly Quasiconvex PAC-Bayesian Bound. *28th Annual Conference on Learning Theory (COLT)*, 2017. 2, 3, 15

[52] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 9

[53] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32, 2019. 9

[54] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997. 8

[55] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the Noisy Gradient Descent that Generalizes as SGD. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. 9

[56] Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets. *The 7th International Conference on Learning Representations (ICLR)*, 2019. 5

[57] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 9

[58] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64 (3):107–115, 2021. 1

[59] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-Vacuous Generalization Bounds at the ImageNet Scale: A PAC-Bayesian Compression Approach. *7th International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 3, 4, 6

[60] Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. *Advances in Neural Information Processing Systems*, 34, 2021. 9

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] We provide evidence for the claims made in the abstract and introduction in Section 4, Section 5.1, and Section 6.
   (b) Did you describe the limitations of your work? [Yes] We describe limitations in Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (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? [Yes] See Appendix I.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix I.
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]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix.
   (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 D.
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]
   (b) Did you mention the license of the assets? [Yes] See Appendix J.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Our code is submitted in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
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]