INTRINSIC DIMENSIONALITY EXPLAINS THE EFFECTIVENESS OF LANGUAGE MODEL FINE-TUNING

Armen Aghajanyan, Luke Zettlemoyer, Sonal Gupta
Facebook
{armenag,lsz,sonalgupta}@fb.com

ABSTRACT

Although pretrained language models can be fine-tuned to produce state-of-the-art results for a very wide range of language understanding tasks, the dynamics of this process are not well understood, especially in the low data regime. Why can we use relatively vanilla gradient descent algorithms (e.g., without strong regularization) to tune a model with hundreds of millions of parameters on datasets with only hundreds or thousands of labeled examples? In this paper, we argue that analyzing fine-tuning through the lens of intrinsic dimension provides us with empirical and theoretical intuitions to explain this remarkable phenomenon. We empirically show that common pre-trained models have a very low intrinsic dimension; in other words, there exists a low dimension reparameterization that is as effective for fine-tuning as the full parameter space. For example, by optimizing only 200 trainable parameters randomly projected back into the full space, we can tune a RoBERTa model to achieve 90% of the full parameter performance levels on MRPC. Furthermore, we empirically show that pre-training implicitly minimizes intrinsic dimension and, perhaps surprisingly, larger models tend to have lower intrinsic dimension after a fixed number of pre-training updates, at least in part explaining their extreme effectiveness. Lastly, we connect intrinsic dimensionality with low dimensional task representations and compression based generalization bounds to provide intrinsic-dimension-based generalization bounds that are independent of the full parameter count.

1 INTRODUCTION

Pre-trained language models (Radford et al., 2019; Devlin et al., 2018; Liu et al., 2019; Lewis et al., 2019; 2020) provide the defacto initialization for modeling most existing NLP tasks. However, the process of fine-tuning them on often very small target task datasets remains somewhat mysterious. Why can we use relatively vanilla gradient descent algorithms (e.g., without strong regularization) to tune a model with hundreds of millions of parameters on datasets with only hundreds or thousands of labeled examples?

We propose intrinsic dimensionality as a new lens through which fine-tuning can be analyzed (Li et al., 2018). An objective function's intrinsic dimensionality describes the minimum dimension needed to solve the optimization problem it defines to some precision level. In the context of pre-trained language models, measuring intrinsic dimensional will tell us how many free parameters are required to closely approximate the optimization problem that is solved while fine-tuning for each end task. For example, we will show that 200 parameters (randomly projected back into the full parameter space) are enough to represent the problem of tuning a RoBERTa model to within 90% of the performance of the full model. More generally, we also describe a set of strong empirical and theoretical connections between intrinsic dimensionality, number of parameters, pre-training, and generalization.

We first empirically show that standard pre-trained models can learn a large set of NLP tasks with very few parameters and that the process of pre-training itself implicitly minimizes the intrinsic dimension of later tuning for different NLP tasks. We continue by conducting a study across over a dozen various pre-trained models to show that number of parameters strongly inversely correlates with intrinsic dimensionality, at least in part to justify the extreme effectiveness of such models. We

interpret pre-training as providing a framework that learns how to compress the average NLP task. Finally, we connect intrinsic dimensional with low dimensional task representations and compression based generalization bounds to provide intrinsic-dimension-based generalization bounds that are independent of the full parameter count, further justifying why these methods generalize so well in practice across tasks.

The contributions of our paper are the following:

- We empirically show that common NLP tasks within the context of pre-trained representations have an intrinsic dimension several orders of magnitudes less than the full parameterization.
- We propose a new interpretation of intrinsic dimension as the downstream fine-tuning task's minimal description length within the framework of the pre-trained model. Within this interpretation, we empirically show that the process of pre-training implicitly optimizes the description length over the average of NLP tasks, without having direct access to those same tasks.
- We measure the intrinsic dimension of a large set of recently developed pre-training methods. We discover that there exists a fortuitous trend where larger models tend to have a smaller intrinsic dimension.
- Lastly, we show that compression based generalization bounds can be applied to our intrinsic dimension framework to provide generalization bounds for large pre-trained models independent of the pre-trained model parameter count.

2 Related Work

Calculating the intrinsic dimension of an objective function was proposed Li et al. (2018). In their paper, they analyzed the impact of various architectures on the intrinsic dimensionality of their objective. Our work is a direct extension of this paper, focusing on analyzing pre-trained representations instead.

There is a large collection of literature analyzing pre-trained models from the perspective of capacity. For example, a recent line of work has shown that pre-trained models such as BERT are redundant in their capacity, allowing for significant sparsification without much degradation in end metrics (Chen et al., 2020; Prasanna et al., 2020; Desai et al., 2019). Houlsby et al. (2019) showed that fine-tuning top layers of pre-trained models is not effective and that alternate methods allow fine-tuning effectively with a couple of percent of the parameters. Furthermore, we can view computing the intrinsic dimensionality as a continuous relaxation of the sparsification problem.

Moreover, standard approaches towards fine-tuning seem to have non-trivial effects on the generalization of pre-trained representations (Aghajanyan et al., 2020). A holistic explanatory picture of the successes of fine-tuning has not yet been painted. A clear understanding of the underlying mechanisms which lead to the incredible generalization of fine-tuned pre-trained representations is currently missing. Moreover, we still do not understand why various pre-training methodology manifests in universally useful representations.

3 Intrinsic Dimensionality of Finetuning

Background An objective function's intrinsic dimension measures the minimum number of parameters needed to reach satisfactory solutions to the respective objective (Li et al., 2018). Alternatively, the intrinsic dimension represents the lowest dimensional subspace in which one can optimize the original objective function to within a certain level of approximation error. Computing the exact intrinsic dimensional of the objective function is computation intractable; therefore, we resort to heuristic methods to calculate an upper bound. Let $\theta^D = [\theta_0, \theta_1, ..., \theta_m]$ be a set of D parameters that parameterize some model $f(\cdot, \theta)$. Instead of optimizing the empirical loss in the original parameterization (θ^D), the subspace method fine-tunes the model via the following re-parametrization in the lower-dimensionsal d-dimensions:

$$\theta^D = \theta_0^D + P(\theta^d) \tag{1}$$

where $P : \mathbb{R}^d \to \mathbb{R}^D$ projects from a parameter from a lower dimensional d to the higher dimensional D. Intuitively, we do an arbitrary random projection onto a much smaller space; usually, a linear projection, we then solve the optimization problem in that smaller subspace. If we reach a satisfactory solution, we say the dimensionality of that subspace is the intrinsic dimension. This methodology was proposed in the seminal paper by Li et al. (2018). Concretely Li et al. (2018) proposed 3 various actualizations of P; a random linear dense projection ($\theta^d W$), random linear sparse projection($\theta^d W_{\text{sparse}}$) and random linear projection via the Fastfood transform (Le et al., 2013).

We will primarily use the Fastfood transform, defined as:

$$\theta^D = \theta_0^D + \theta^d M \qquad\qquad M = HG\Pi HB \tag{2}$$

The factorization of M consists of H, a Hadamard matrix, G, a random diagonal matrix with independent standard normal entries, B a random diagonal matrix with equal probability ± 1 entries, and Π a random permutation matrix. Furthermore, the matrix multiplication with a Hadamard matrix can be computed in $\mathcal{O}(D \log d)$ via the Fast Walsh-Hadamard Transform. Note that everything but θ_d is fixed; therefore, the optimization problem lies only in d-dimensions. Note that if we place a constraint of M being a binary matrix, we recover the sparsification problem; therefore, we can view finding intrinsic dimensionality as a continuous relaxation of the sparsification problem.

The standard method of measuring the intrinsic dimensionality of an objective as proposed by Li et al. (2018) requires searching over various d, training using standard SGD over the subspace reparameterization θ^D and selecting the smallest d which provides us with a satisfactory solution (d_{90}). Li et al. (2018) defined the *satisfactory solution* as being 90% of the full training metric. For example, if we reach 85% accuracy training a model with all of its parameters, the goal is to find the smallest d, which would reach $0.9 * 85\% = 76.5\%$ accuracy; we call this dimension d_{90}. Let us also note that by merely initializing $\theta^d = 0$ we recover the original parameterization θ_0^D which in the context of fine-tuning represents the original weights of the pre-trained model.

The way Li et al. (2018) define a satisfactory solution reduces the dependence of the dataset's size on the calculation of intrinsic dimension. For a small dataset, we will generally have worse end metrics; therefore, we have a lower d_{90} cut-off; inversely, a larger dataset will require a more non-trivial d_{90} cut-off.

Structure Aware Intrinsic Dimension Due to the large size of pre-trained language models (generally in the hundreds of millions of parameters), the only computationally reasonable subspace optimization method is one that utilizes the Fastfood transform. For example, if we are interested in subspace training with $d = 1000$ for the RoBERTa-Large model using a dense matrix, we would require 1.42 terabytes of memory to store just the projection matrix.

Unfortunately, the method of finding the intrinsic dimension proposed by Li et al. (2018) is unaware of the layer-wise structure of the function parameterized by θ. Existing literature argues that in attention-based pre-trained models, individual layers specialize separately (Clark et al., 2019); therefore, it is useful to incorporate a notion of structure when computing d_{90}. We define Structure-Aware Intrinsic Dimension (SAID) as the following

$$\theta_i^D = \theta_{0,i}^D + \lambda_i P(\theta^{d-m})_i \tag{3}$$

For m layers, we trade m parameters from our subspace parameter θ_d to allow for layer-wise scaling through jointly learned λ, thus θ_d becomes $[\theta_{d-m}, \lambda]$. This allows the SAID method to focus a larger capacity of θ^{d-m} towards specific layers what might carry more relevant information for the task at hand. Conversely, we will refer to the layer unaware method (Equation 2) as the Direct Intrinsic Dimension (DID) method.

4 Intrinsic Dimensionality of Common NLP Tasks

4.1 Sentence Prediction

We first empirically calculate the intrinsic dimension of various pre-trained models on a set of sentence prediction tasks from the GLUE Benchmark (Wang et al., 2018). We focus on analyzing BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) at both the base and large model sizes.