Figure 1: The following figures show the evaluation accuracy on two datasets and four models across a range of dimensions d for the DID method. The horizontal lines in each figure represent the 90% solution of the respective full model.

We chose to experiment with MRPC (Dolan & Brockett, 2005) and QQP (Iyer et al., 2017) as reference examples of small and large tuning datasets. MRPC is a binary classification task for predicting semantic equivalency for two paraphrases with roughly 3700 training samples, while QQP is a binary classification task for predicting semantic equality of two questions, with roughly 363k samples. For every dataset and every model, we run 100 subspace trainings with d ranging from 10 to 10000 on a log scale. For every training run, we do a small hyperparameter search across four learning rates. We initialize every θ_d to the zero vector to allow for our starting point to be the original pre-trained model. Our subspace optimization method also operates over the randomly initialized sentence classification head to ensure we have exactly d parameters to optimize.

We use both the SAID and DID subspace optimization methods, which we implemented in the Huggingface Transformers library (Wolf et al., 2019). We present the results in Figure 1.

4.2 Analysis

The first takeaway is the incredible low dimensionality of viable solutions. With RoBERTa-Large, we can reach 90% of the full fine-tuning solution of MRPC using roughly 200 parameters and 800 parameters for QQP (Table 1). Recall that our approximation of intrinsic dimension is necessarily crude by using random projections and restricting them to the use of Fastfood transform; therefore, it is likely that the true intrinsic dimension is much lower.

	SAID		DID	
Model	MRPC	QQP	MRPC	QQP
BERT-Base	1608	8030	1861	9295
BERT-Large	1037	1200	2493	1389
RoBERTa-Base	896	896	1000	1389
RoBERTa-Large	**207**	**774**	322	**774**

Table 1: Estimated d_{90} intrinsic dimension for a set of sentence prediction tasks and common pre-trained models. We present both the *SAID* and *DID* methods.

Furthermore, RoBERTa consistently outperforms BERT across various subspace dimensions d while having more parameters. We leave a more in-depth analysis of model parameter size on intrinsic dimensionality to a later section (§5.2).

Lastly we see that adding a notion of structure in the computation of intrinsic dimension is beneficial with the SAID method consistently improving over the structure unaware DID method.

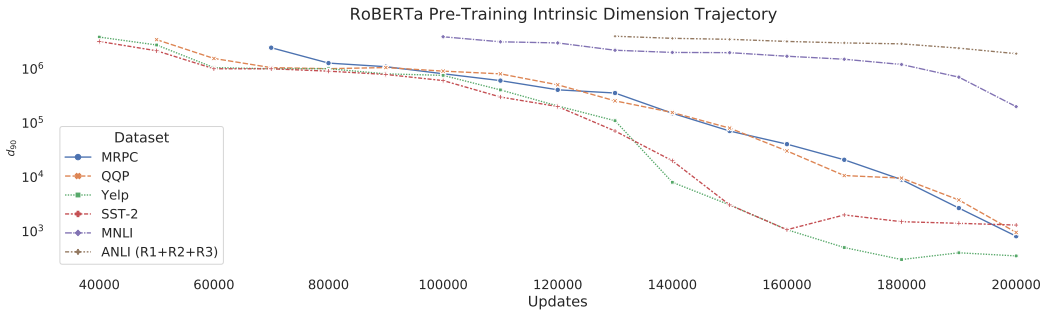

Figure 2: Every 10k updates of RoBERTa-Base that we trained from scratch, we compute d_{90} for six datasets; MRPC, QQP, Yelp Polarity, SST-2, MNLI, and ANLI. If we were unable to compute a d_{90} for a specific checkpoint, we do not plot the point, hence some datasets start at later points. Unable to compute means either we could not fine-tune the full checkpoint to accuracy above majority class or stabilize SAID training.

5 Intrinsic Dimension, Pre-Training, and Generalization Gap

One interpretation of the intrinsic parameter vector is that it encodes the task at hand with respect to the original pre-trained representations. Therefore, we can interpret d as the minimal description length of the task within the framework dictated by the pre-trained representations (Hinton & Zemel, 1993). Under this interpretation of intrinsic dimensionality, we hypothesize that pre-training is implicitly lowering the intrinsic dimensionality of the average NLP task, and therefore compress the minimal description length of those same tasks.

What do we more precisely mean by intrinsic parameter encoding a task within the framework provided by the pre-trained representations? Traditionally, a finetuned model (e.g. for a classification tasks) simply consists of a classification head g, parameterized by w_g applied to fine-tuned representations f, parameterized by w_f per sample x. Therefore, to fully describe a task, we need to pack together parameterizations and weights $\{g, f, w_g, w_f\}$. This model description is completely decoupled from the original weights of the pre-trained representation w_{f_0}, therefore to represent n classification tasks, we need to maintain $n\{w_g, w_f\}$; additionally, the task representation is incredibly high dimensional. Conversely, fine-tuning utilizing SAID in d-dimensions requires storing only θ_d per task, a single random seed used to generate M and the original pre-trained weights w_{f_0}. Therefore, we can represent arbitrary NLP tasks within a single pre-trained model framework with $d + 1$ parameters.

For example, in the last section, we represented MRPC with roughly 200 parameters, which translates to needing less than a kilobyte of data to encode a complex natural language task within the framework provided by RoBERTa.

We hypothesize that the better the pre-trained models are, the fewer bits (description length) are needed to represent the average NLP task, as we will demonstrate empirically in the next section.

5.1 Pre-Training Intrinsic Dimension Trajectory

To verify our hypothesis of pre-training optimizing intrinsic dimension, we retrain a RoBERTa-Base from scratch and measure various NLP tasks' intrinsic dimensions using the SAID method across various checkpoints. We completely replicate the setting as described by (Liu et al., 2019) apart from only training for a total of 200k steps (instead of 500k) with half the batch size (1k). To calculate the intrinsic dimension more efficiently, we reuse the best learning rates discovered in Section 4 for $d < 10000$ and use a fixed learning rate for anything else. To find d_{90} we do a binary search across d per each checkpoint, with a minimum d of 100 and a maximum of 4 million. The "full solution" that we use when deciding d_{90} cut-off is computed by fine-tuning the checkpointed model in the standard way. We compute SAID on six datasets; *MRPC*, *QQP*, *Yelp Polarity* (Zhang et al., 2015), *SST-2* (Socher et al., 2013), *MNLI* (Williams et al., 2018) and *ANLI* using all rounds of data (Nie et al., 2019).

We present our results in Figure 2. We see that the intrinsic dimensionality of RoBERTa-Base monotonically decreases as we continue pre-training. We do not explicitly optimize for intrinsic dimensionality, specifically during pre-training (the language model does not have access to downstream datasets!), but none-the-less the intrinsic dimension of these downstream tasks continues to decrease.

More so, tasks that are easier to solve consistently show lower intrinsic dimensionality across all checkpoints, for example, *Yelp Polarity* vs. the notoriously tough *ANLI* dataset. The correlation between tasks traditionally hard for RoBERTa and their large intrinsic dimension hints at a connection between generalization and intrinsic dimension. We will discuss generalization further in Section §5.3.

Given our task representation interpretation of intrinsic dimensionality, we argue that the large scale training of Masked Language Models (MLM) learns generic and distributed enough representations of language to facilitate downstream learning of highly compressed task representations. Furthermore, we argue for another perspective of pre-training learning representations that form a compression framework with respect to various NLP tasks.

5.2 Parameter Count and Intrinsic Dimension

We would also like to measure the relationships between the parameter count of arbitrary pre-trained models and the intrinsic dimension of downstream NLP tasks. The optimal experiment to run would be to fix the pre-training method, e.g., MLM RoBERTa style, vary the architecture size from small to very big, and compute the intrinsic dimension of a group of tasks at every size of the model. Unfortunately, such an experiment is computationally infeasible due to the need to train many RoBERTa models.

Due to these constraints, we opt to do an empirical study over existing pre-trained models, regardless of the pre-training method. We show that the trend is strong enough to overcome differences in training methodology. We select the following pre-trained models in our study: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2019), Electra (Clark et al., 2020), Albert (Lan et al., 2019), XLNet (Yang et al., 2019), T5 (Raffel et al., 2019), and XLM-R (Conneau et al., 2019). Furthermore, we selected various sizes of these models, as available publicly within the HuggingFace Transformers library (Wolf et al., 2019).

We used the MRPC dataset and computed intrinsic dimension for every pre-trained model utilizing the same binary search methodology mentioned in the previous section with additional small hyperparameter searches across learning rate (due to the wide range of learning rates needed by various models).

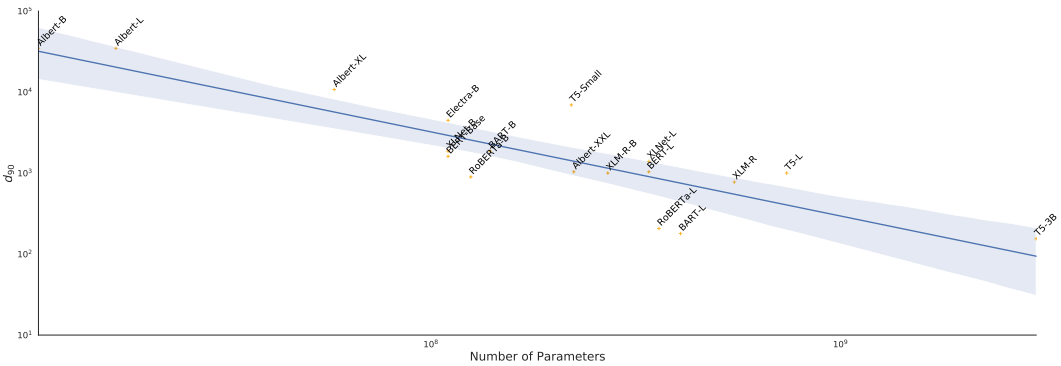

Figure 3: We calculate the intrinsic dimension for a large set of pre-trained models using the SAID method on the MRPC dataset.

We present our results in Figure 3. We see a strong general trend that as the number of parameters increases, the intrinsic dimension of fine-tuning on MRPC decreases. We ran this experiment on other datasets to ensure that this is not an artifact of the dataset. Our experiments showed the same trend; we refer to the Appendix for all trends per dataset.