

Figure 4: We plot the evaluation accuracy of six datasets across various intrinsic dimensionalities. There is a strong general trend that pre-trained models that are able to attain lower intrinsic dimensions generalize better.

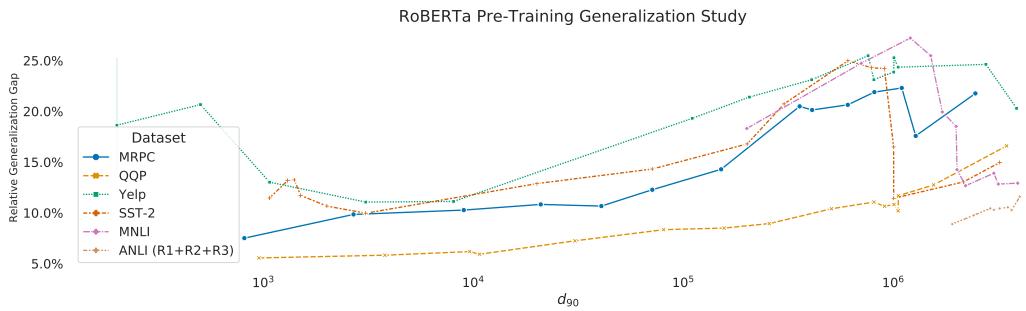


Figure 5: We plot the intrinsic dimension and the respective relative generalization gap across a set of varied tasks.

Within the same window of number of parameters, pre-training methodology becomes essential. For example, in the regime of 10^8 parameters, the RoBERTa method of pre-training dominates similar sized pre-training methods. However, there does not seem to be a method that can overcome the limitations induced by the number of parameters. Interpreting these results through the lens of learning a compression framework for NLP tasks is straightforward; the more parameters we have in the model, the less we need to represent a task.

5.3 GENERALIZATION BOUNDS THROUGH INTRINSIC DIMENSION

We have shown strong empirical evidence connecting pre-training, fine-tuning, and intrinsic dimensionality. However, we have yet to argue the connection between intrinsic dimensionality and generalization. Given that we have seen pre-training minimize intrinsic dimension, we hypothesize that generalization improves as the intrinsic dimension decreases.

To do so, we will empirically experiment with the connections between d_{90} and evaluation set performance by looking at various checkpoints from our RoBERTa experiments in Section §5.1. We also plot the relative generalization gap (delta between train time performance and test time performance).

In Figure 4 we plot the evaluation accuracy’s achieved by our pre-training experiment in Section §5.1. A lower intrinsic dimension is strongly correlated with better evaluation performance. Additionally we are interested in measuring relative generalization gap ($\frac{acc_{train} - acc_{eval}}{1 - acc_{eval}}$) across intrinsic dimension. We select the training accuracy that provides us with the best evaluation metrics when computing this figure.

We present our results in Figure 5. Lower intrinsic dimension once again correlates strongly with a smaller relative generalization gap. If we interpret the intrinsic dimension as a measure of complexity, we expect the generalization gap to decrease with intrinsic dimension.

5.3.1 GENERALIZATION BOUNDS

By applying standard compression based generalization bounds, we can provide theoretical backing to the empirical connection between intrinsic dimension and generalization (Arora et al., 2018).

Consider the following definition of multi-class classification loss with an optional margin over our supervised dataset D .

$$\mathcal{L}_\gamma(f) = \mathbb{P}_{(x,y) \sim D} \left[f(x)[y] \leq \gamma + \max_{i \neq y} f(x)[j] \right] \quad (4)$$

When $\gamma = 0$, \mathcal{L}_0 recovers the standard classification loss. Furthermore, Let $\hat{\mathcal{L}}_\gamma(f)$ be an unbiased empirical estimate of the margin loss.

Theorem 1. *Let f be a function which is parameterized by θ^D as described in Equation 1 with a total of d trainable intrinsic parameters on a dataset with m samples. Then with a high probability, we can state the following asymptotic generalization bound*

$$\mathcal{L}_0(f) \leq \hat{\mathcal{L}}_0(f) + \mathcal{O}\left(\sqrt{\frac{d}{m}}\right) \quad (5)$$

Proof. We defer the proof Section §A.1 in the Appendix. We note that this is an extension of the well-known compression based generalization bound explored by Arora et al. (2018). \square

This generalization bound is independent of the underlying parameter count (D) of the pre-trained model but depends on the ability to compress the downstream task (d). Moreover, given that our previous section shows larger models compress better, our bounds are aligned with general intuition and recent empirical evidence that larger pre-trained models generalize better. Explicitly, these bounds only apply to pre-trained methods trained with the intrinsic dimension subspace method; research has yet to show that standard SGD optimizes in this low dimensional space (although experimentally, this seems to be confirmed). We leave the theoretical contribution of showing SGD optimizes in this space, resembling something such as intrinsic subspace, for future work.

We want to highlight that generalization is not necessarily measured by the pre-trained model's parameter count or measure of complexity, but the pre-trained model's ability to facilitate the compression of downstream tasks. In some sense, if we want to compress downstream tasks better, we must expect pre-trained representations to have a considerable measure of complexity.

6 CONCLUSION

In conclusion, we proposed viewing the various phenomena surrounding fine-tuning and pre-training through the lens of intrinsic dimensionality. We empirically showed that common natural language tasks could be learned with very few parameters, sometimes in the order of hundreds, when utilizing pre-trained representations. We provided an interpretation of pre-training as providing a compression framework for minimizing the average description length of natural language tasks and showed that pre-training implicitly minimizes this average description length.

We continued by doing an empirical study of existing pre-training methods and their respective intrinsic dimension, uncovering the phenomena that intrinsic dimensionality decreases as we increase the number of pre-trained representation parameters. This phenomenon provides some intuitions to the trend of growing pre-trained representations. We connected intrinsic dimensionality with generalization by first showing that pre-trained models with lower intrinsic dimensions across various tasks achieve higher evaluation accuracies and lower relative generalization gaps. Furthermore, we explain these empirical results by applying well-known generalization bounds to the intrinsic dimension to get generalization bounds that grow on the order of the intrinsic dimension, not on the pre-trained model's parameter count.

Intrinsic dimensionality is a useful tool for understanding the complex behavior of large models. We hope that future work will make explicit theoretical connections between SGD and optimizing the intrinsic dimension as well as explain exactly why pre-training methods optimize the intrinsic dimensionality of tasks before not seen.

REFERENCES

- Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*, 2020.
- Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.
- Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained bert networks. *Advances in Neural Information Processing Systems*, 33, 2020.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert’s attention. *arXiv preprint arXiv:1906.04341*, 2019.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
- Shrey Desai, Hongyuan Zhan, and Ahmed Aly. Evaluating lottery tickets under distributional shifts. *arXiv preprint arXiv:1910.12708*, 2019.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.
- Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6:3–10, 1993.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*, 2019.
- Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. First quora dataset release: Question pairs, 2017. URL <https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs>.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
- Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, volume 85, 2013.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
- Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. Pre-training via paraphrasing, 2020.
- Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.