Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. When bert plays the lottery, all tickets are winning. *arXiv preprint arXiv:2005.00561*, 2020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://www.aclweb.org/anthology/W18-5446`.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/N18-1101`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pp. arXiv–1910, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pp. 5753–5763, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. *arXiv:1509.01626 [cs]*, September 2015.

A Appendix

A.1 Proofs

Arora et al. (2018) define (γ, S) compressible using helper string s as the following.

Definition 1. *(γ, S) compressible using helper string s*

Suppose $G_{\mathcal{A},s} = \{g_{\theta,s} | \theta \in \mathcal{A}\}$ is a class of classifiers indexed by trainable parameters A and fixed strings s. A classifier f is (γ, S)-compressible with respect to $G_{\mathcal{A}}$ using helper string s if there exists $\theta \in \mathcal{A}$ such that for any $x \in S$, we have for all y

$$|f(x)[y] - g_{\theta,s}(x)[y]| \leq \gamma \tag{6}$$

Remark 1. *If we parameterize $f(x;\theta)$ via the intrinsic dimension approach as defined in Equation 1, then f is compressible losslessly using a helper string consisting of the random seed used to generate the static random projection weights and the initial pre-trained representation θ_0^D. Therefore we say f parameterized by either DID or SAID is $(0, S)$ compressible.*

Theorem 2.1 in Arora et al. (2018) states given a compression consisting of r discrete states we achieve the following generalization bound.

$$\mathcal{L}_0(f) \leq \hat{\mathcal{L}}_\gamma(f) + O\left(\sqrt{\frac{d \log r}{m}}\right) \tag{7}$$

We can trivially represent our parameters θ_d in a discrete fashion through discretization (as was done in Arora et al. (2018)), and the number of states is dependent on the level of quantization but is static once chosen (FP32 vs. FP16).

We then connect the fact that models trained in low dimensional subspace using SAID/DID methods are (0, S)-compressible to derive the final asymptotic bound.

$$\mathcal{L}_0(f) \leq \hat{\mathcal{L}}_0(f) + \mathcal{O}\left(\sqrt{\frac{d}{m}}\right) \tag{8}$$