Relating pretrain dataset diversity to BERT's finetuned performance

Jon Ball
Stanford Graduate School of Education
jonball@stanford.edu

Abstract

Additionally pretraining neural language models on domain-specific data can improve task performance. But selecting these data is a process usually determined by data availability in a given domain. Assuming that relevant pretrain documents are sufficiently abundant, how can researchers select the documents that most contribute to a model's subsequent, finetuned performance? **Diversity** is one possible criterion for selecting pretrain documents. However, preliminary results do not support the claim that additional pretraining on a maximally diverse dataset improves model performance relative to pretraining on a random dataset, holding the model and domain constant. Pretraining a BERT base model on additional datasets of 10,000 news articles each, wherein each article is selected from among more than 700,000 competing articles on the basis of a formal diversity score, does not significantly impact model performance on the SQuAD Natural Language Understanding task (v1.1). Future research may yet establish that pretrain dataset diversity contributes to finetuned model performance, although a successful experiment will require pretraining on diverse datasets consisting of many more than 10,000 examples each.

1 Introduction

The operative paradigm for training large language models, or foundation models (Bommasani et al., 2021), involves maximizing the quantity of pretrain data as well as model size in terms of parameters. This quantity-over-quality paradigm has a firm basis in observed model performance. (Gordon et al., 2021) Prioritizing data "quality" or other input characteristics would be justifiable in the event that research establishes a robust connection between these characteristics and model performance, holding the model size and number of examples constant. This is an admittedly narrow justification that derives from the narrowness of evaluative tasks. Given the critical media reception of problematic chatbots and other language model applications, an up-front investment in data quality would seemingly pay dividends. In any event, such would naturally require the adoption of a set of metrics for data quality, or in the case of this article, diversity. The diversity metrics examined below are drawn from corpus linguistics (mean-IDF), pre-neural language modeling (trigram language model entropy), and embedding spaces ("Outlier Diversity" and "Word Embedding Diversity").

It bears stating explicitly at the outset that quality and diversity are not synonymous. Prior research provides a theoretical explanation for why diversity *might* approximate quality with regard to a language model's pretrain data inputs. Aghajanyan et al. (2021) offered an explanation for why finetuning a model on task-relevant data subsequently improves its performance, which relies on the notion of the "intrinsic dimensionality" of tasks. The authors argued that language models adapt to a task's intrinsic dimensionality while finetuning on task-specific data. The theory in support of data diversity follows: **If more diverse inputs encourage a model to learn more or different data features on which it can fit during the finetuning process, then diversity may improve a model's ability to later model and solve a task-specific objective function.** If diverse data do in fact provide a better opportunity for a model to learn, then this opportunity would seemingly consist of providing additional or more varied data features, because the characteristics and parameter count of the model itself are fixed.

Posing the question at a high level: How does the lexical or semantic diversity of a pretrain dataset relate to a language model's finetuned task performance? Limited evidence suggests that pretraining on more diverse data may improve downstream performance on certain tasks, such as emotion classification and dialogue generation. (Stasaski et al., 2020) This article extends the insight that diverse pretrain data may be useful to a well-known Natural Language Understanding (NLU) task: the Stanford Question Answering Dataset (SQuAD v1.1). (Rajpurkar et al., 2016) However, evidence also shows that domain- and task-adaptive pretraining methods (Gururangan et al., 2020), which in some sense involve limiting diversity, also improve performance. The empirical question of how pretrain dataset characteristics such as diversity relate to a finetuned language model's eventual task performance remains largely unanswered. It is nonetheless intuitive that a model benefits by learning from the "right" data for the task at hand, and that the benefits of data diversity would be most strongly felt if the diverse data in question fit within a primary domain.

2 Prior Literature

Beginning chronologically, Radford et al. (2018) innovated a general, two-phase process for training transformer-based language models. The authors improved on several benchmarks at the time by first pretraining their transformer model on unlabeled text with a *generative* task, and then finetuning on a supervised *discriminative* task. This two-phase approach is by now widely adopted and forms the basis for the experiment reported in this article. Radford et al. (2018)'s findings demonstrate the overall utility of sequencing different pretraining and finetuning tasks. Their findings hint at the possibility that training tasks can be optimally sequenced, depending on the intrinsic characteristics of (a) language models, (b) tasks, and (c) (pre-)training data. Additionally, Radford et al. proposed an explanatory hypothesis which has arguably received some validation since their article was published: "that the underlying language model learns to perform many of the tasks we evaluate on in order to improve its language modeling capability," thereby improving downstream task performance. (Radford et al., 2018, pgs. 7-8)

Expanding beyond Radford et al.'s two-phase approach, Gururangan et al. (2020) demonstrated that "multi-phase adaptive pretraining" improves downstream performance over simpler, two-phase implementations. Gururangan et al. also proposed two complementary pretraining methods: "task-adaptive pretraining" (TAPT) and "domain-adaptive pretraining" (DAPT). The authors did not offer a theoretical explanation for why in-domain and task-specific pretraining result in performance gains. However, they empirically demonstrated that models with millions upon millions of parameters such as RoBERTa nevertheless struggle "to encode the complexity of a single textual domain." (Gururangan et al., 2020, pg. 8350) Greater opportunity to encode this complexity corresponds to better performance, in other words. A possible theoretical explanation for the efficacy of TAPT and DAPT is that pretraining on additional, relevant data encourages a model to focus on subspaces within domains, particularly those which are related to performance on a given task. Extending Radford et al.'s hypothesis, additional pretraining along the lines of TAPT and DAPT may improve downstream task performance because it offers language models additional opportunities to learn how to model language in task-relevant dimensions, which can then be used to model the objective function during finetuning.

Building on this theoretical intuition, Aghajanyan et al. (2021) offered a compelling explanation for why the pretraining-finetuning approach works well in practice. The authors analyzed BERT and RoBERTa with reference to the concept of "intrinsic dimension." (Li et al., 2018) Intrinsic dimension is the lowest dimensional subspace in which the objective function for a given task can be optimized by a given language model to within a reasonable error threshold. Rephrased, Aghajanyan et al. discovered that finetuning allows a language model to focus on the dimensions of variation in a high-dimensional embedding space that are most appropriate for solving a task. The authors also discovered that the number of language model parameters (often in the billions) strongly *inversely* correlates with intrinsic dimensionality, meaning that language models with many more parameters reduce themselves to a much lower dimensionality when finetuning on new tasks. This means that language models which have learned more may be able to reduce tasks to optimization problems in fewer dimensions, although this reduction perhaps contributes to the "forgetting problem." (He et al., 2021) Aghajanyan et al.'s explanation for the effectiveness of finetuning arguably applies to pretraining as well. Domain-adapted language models such as SciBERT may perform well on specific tasks precisely because they do not expend compute learning patterns with little explanatory value in the task domain, albeit at the cost of generalizability. (Beltagy et al., 2019)

If the tasks we assign to language models have an intrinsic dimension, and if language models are better at adapting to this dimensionality the more parameters they have, then another question remains: How can training data be selected to improve the finetuning process of dimensionality reduction? Again, a possible theoretical explanation for the positive effects of pretraining described by Gururangan et al. (2020) is that pretraining with a language modeling objective on domain- or task-specific data causes language models to learn along dimensions intrinsically suited to the task at hand, and even to disregard irrelevant dimensions or data features.

The corresponding, unproven hypothesis is this: **Diverse pretraining data improve a language model's performance on NLP tasks by encouraging the model to learn from a more numerous or varied sample of data features, on which it can then fit while finetuning.** Corollaries of this hypothesis read as caveats. Prioritizing data diversity would have little positive impact on model performance if the diversity metric used did not result in the sampling of additional or more varied features, relative to the sample which would have been selected regardless. Because additional pretrain data, per Gururangan et al.'s example, tend to be domain-specific, it may be the case that relevant data features are learned regardless of how diverse the domain-specific data are. Put simply: the domain, not the diversity within it, is what might actually matter. Furthermore, in its maximal sense, training on diverse data implies training on task-irrelevant data, which may amount to a waste of compute.

As a final caveat, it bears stating that the hypothesis above could be fundamentally flawed because it relies on a specific understanding of how neural language models actually learn. *Local linear explanations* (LLEs) offer an established but flawed framework for peering into the "black box" that is a neural network. (Amparore et al., 2021) Assuming that a neural net does something like fitting local linear models on local data features, which would presumably be more numerous or varied in a more diverse sample of data, could in fact be an erroneous assumption.

To actually test the hypothesis that the diversity of a pretrain dataset positively correlates with a model's finetuned task performance requires a set of metrics for diversity. Palumbo et al. (2020, pg. 47) proposed Word Embedding Distance (WED), defined as "the average cosine distance between embeddings of vectors" in a dataset. Similarly, Stasaski et al. (2020, pgs. 4962-3) proposed an "outlier" measure of corpus diversity, defined as "the Euclidean distance between an utterance embedding and the average embedding for all utterances in the sub-corpus." (Larson et al., 2019) While these metrics are clearly suited for pairwise comparison, the intuition behind them – averaging variation along all the dimensions in an embedding space – has been further developed by Lai et al. (2020). Non-embedding approaches to measuring dataset diversity such as entropy scores determined by n-gram language modeling and Mean-IDF (Baeza-Yates et al., 1999) may also be useful.

3 Data

The additional pretrain data used to test the diversity hypothesis were drawn from a subset of the Common Crawl news corpus. The dataset initially contained 708,241 English language news articles published between Jan 2017 and December 2019. These were accessed using Hugging Face's datasets module and can be previewed at this link. A corresponding jupyter notebook detailing the implementation of diversity metrics and the selection of samples with reproducible code can be viewed at this link.

The evaluation data were simply the Stanford Question Answering Dataset v1.1 (SQuAD v1.1), which can be accessed via Pranav Rajpurkar's dedicated webpage. (Rajpurkar et al., 2016) The f1-score metric was ultimately used to assess model performance on the held out SQuAD test set.

One article included in the CC news sub-corpus is actually written in Gaelic, and another several feature Spanish or sequences of non-English characters. In order to prevent these diverse but not-so-relevant-for-SQuAD articles from being selected on the basis of each diversity metric, articles were filtered out which were not reliably classified as English by Google's Compact Language Detector v3 (CLD3). The resulting dataset consisted of 703,532 articles.

Five separate datasets were ultimately passed as additional pretrain inputs for masked language modeling to five separate BERT base models. Four such datasets consisted of 10,000 articles selected on the basis of four different diversity metrics, and the fifth control dataset consisted of 10,000 articles chosen at random. The datasets contained some shared articles. A qualitative observation of each dataset revealed that sports articles containing strings of numbers corresponding to game scores (e.g., goals, points) yielded high diversity scores, and were consequently included. Relative to the majority of the corpus, which are articles written fully in English and without excessive numeration, sport score reports are indeed "diverse." It is a potential drawback of several of the diversity metrics used that they compare a given document against a corpus average. This privileges relatively diverse articles such as numeric sport reports despite their not being intrinsically diverse. Lai et al. (2020)'s diversity metric, computed as the geometric average of standard deviations along each axis in an embedding space, potentially for a single document, may mitigate this problem. However, Lai et al. (2020)'s metric was not implemented in this research project and deserves to be tested in future research relating train dataset characteristics to model performance.

4 Model

A data card for the single BERT-base-uncased model deployed in this research project (Devlin et al., 2019) can be accessed on the Hugging Face Hub at this link. The most basic BERT model was chosen for its accessibility and high profile. The experiment described below yielded null results. But if it were to have been the case that BERT, the quintessential NLP transformer model, achieved a higher f1-score on SQuAD v1.1, a quintessential NLU task, as a result of having additionally pretrained on a diverse selection of data relative to a random selection of data, then such might have constituted a strong finding in favor of the value of pretrain dataset diversity.

5 Methods

Five separately initialized BERT base uncased models, all sharing the same model parameters, were additionally pretrained on five different datasets consisting of 10,000 English-language Common Crawl news articles each. **The differing composition (i.e., varying diversity) of the pretrain datasets provided the only source of exogenous variation in the research design.** After pretraining each BERT model on 10,000 additional articles for two epochs with a masked language modeling (MLM) objective, the updated model parameters were saved. The five additionally pretrained BERT models were then finetuned and evaluated on the SQuAD v1.1 dataset, using the default specifications provided by this script. Random seeds were preset to ensure reproducible results.

The first additional pretrain dataset consisted of 10,000 articles chosen at random. The BERT model additionally pretrained on this random dataset provided the experiment's control comparison.

The second dataset consisted of 10,000 articles selected from among 703,532 competitors on the basis of their Mean Inverse Document Frequency (mean-IDF). For each article art_i in the sub-corpus sub_c, and for all tokens in the article $t_{i,j}...t_{i,k}$, a mean-IDF score was calculated as:

$$\frac{1}{|\{art\}|} \sum_{t \in art} \log_e \left(\frac{|\{sub_c\}|}{|\{art \,|\, t \in art\}|} \right) \tag{1}$$

where $\{sub_c\}$ is the set of all articles and $\{art|t \in art\}$ is the set of all articles featuring token t_j. (Stasaski et al., 2020)

The articles comprising the third dataset were selected on the basis of trigram language model entropy, which was calculated for a given article as:

$$-\frac{1}{|x \in Trigram(art)|} \sum_{\substack{x \in \\ Trigram(art)}} p(x) \log p(x) \tag{2}$$

The fourth set of articles were selected by first averaging the BERT word embeddings, element-wise, for all tokens $t_j...t_k$ in a given article art. Then the article embedding, E_{art}, was compared against a corpus mean vector, E_{sub_c}, using a simple Euclidean distance metric that produced a single diversity score for each article. Larson et al. (2019) and Stasaski et al. (2020) call this an "Outlier" measure of diversity:

$$E_{art} = \frac{1}{|\{art\}|} \sum_{\substack{t \in \\ art}} {}_{i=1}^{n} BERT(t) \tag{3a}$$

$$E_{sub_c} = \frac{1}{|\{sub_c\}|} \sum_{\substack{art \in \\ sub_c}} {}_{i=1}^{n} E_{art} \tag{3b}$$

$$\sqrt{\sum_i (E_{art} - E_{sub_c})^2} \tag{3c}$$

Articles in the fifth and final dataset were selected from the sub-corpus to maximize the dataset's average "Word Embedding Diversity (WED)" score. (Palumbo et al., 2020) For each article, a single embedding was computed using the same approach as outlined above in (3). Then, the WED score for a given article art_i was computed as the average cosine distance between that article's embedding, E_{art_i}, and the embeddings for all *other* articles in the sub-corpus, sub_c. For ease of notation, the number of articles comprising sub_c is simply defined as N below:

$$WED_{art_i} = \frac{1}{N-1} \sum_{\substack{art \in sub_c \\ \& \, art \neq art_i}} 1 - cos(E_{art_i}, E_{art}) \tag{4}$$

6 Results

The diversity hypothesis presented in this paper would have received preliminary support had any of the four BERT models additionally pretrained on the non-random datasets, comprising articles selected on the basis of the metrics just described, exhibited *significantly* better performance on SQuAD than the random control model. This was **not** the case. In fact, none of the models achieved a significantly higher f1-score than the BERT base model. The SQuAD results achieved by each additionally pretrained model are reported below:

Model	Exact	f1
Random	80.80	88.23
Mean-IDF	80.83	88.25
Entropy	80.96	88.36
Outlier	80.77	88.23
WED	80.67	88.17

(5)

For reference, Devlin et al. (2019) reported an exact match score of 80.8 and an f1-score of 88.5 when they finetuned and evaluated BERT's initial incarnation on SQuAD v1.1. The results presented above do not vary significantly from BERT's baseline performance on SQuAD. At the risk of overinterpreting null results, one tentative conclusion may be drawn. The set of pretrain articles selected on the basis of trigram language model entropy corresponded to relatively better SQuAD performance. Caveats abound, but it makes intuitive sense that another, simpler language model would be well suited for selecting additional pretrain data.

7 Analysis

The most direct conclusion to draw from these results would be that more than 10,000 pretrain examples are required to significantly improve BERT's performance on SQuAD. While the datasets used in this research project varied slightly in terms of average article length, they contained at most 5M words each. This pales in comparison to the size of the subset of the TriviaQA dataset (Joshi et al., 2017) originally used by Devlin et al. (2019) as additional pretrain data for SQuAD, which comprised more than 200M words.[1] Devlin et al. (2019) report that additionally finetuning BERT on TriviaQA led to a modest improvement in SQuAD f1-score of 0.1-0.4.

The quantity-focused paradigm of language model training still applies to quality-selected data. Even the highest quality pretrain datasets, whether their quality is measured as diversity, density, homogeneity, etc. (Lai et al., 2020), must be sufficiently large to impact model performance. Pretraining especially requires an ample quantity of data to avoid overfitting prior to finetuning. Furthermore, with regard to pretraining, readers should again consider that the additional pretraining protocol described in this article consisted solely of two epochs with a masked language modeling (MLM) objective, and *not* any additional pretraining with a next sentence prediction (NSP) objective. This was solely due to constraints on time and compute.

It may well be the case that the Common Crawl news articles comprising the sub-corpus, all of which were published from 2017 to 2019, did not contain information relevant to SQuAD v1.1, which was published in 2016. CC news was nevertheless chosen as a sub-corpus because it seemed plausible that a diverse selection of articles would yield a diverse array of data features, which is an important condition in the diversity hypothesis as earlier stated. If TriviaQA had instead been chosen as the sub-corpus, for example, it seemed likely that a diversity metric could have haphazardly enabled the selection of the "right" QA examples to succeed narrowly on SQuAD. This would provide spurious support for the value of dataset diversity. This line of reasoning also applies to other QA datasets such as HotpotQA (Yang et al., 2018). Relatedly, it may be the case that SQuAD was an inappropriate NLU task to test the value of pretrain dataset diversity. Another task such as emotion classification, which

[1] Devlin et al. report having only pretrained on the first 400 tokens of each evidence document in TriviaQA. If they pretrained on all 660,000+ question-answer-evidence triples, their additional pretrain dataset can be estimated as having been two orders of magnitude larger than the datasets used in this research project, in terms of words.

is the task Stasaski et al. (2020) chose, may provide a better opportunity for a language model to fit on the diverse data features it learns from pretraining.

Finally, although BERT is a well-known if not the *best*-known language model, Talmor et al. (2020) argue that RoBERTa exhibits qualitatively superior performance on reasoning tasks relative to BERT. On the one hand, the research design outlined in this article held model type constant, such that the choice of specific transformer model would hypothetically have no bearing on the results. On the other hand, Talmor et al. (2020)'s point that different models exhibit qualitatively different behavior is well taken. Perhaps RoBERTa would better learn and apply insights from diverse data.

8 Conclusion

This research project failed to find empirical support for the hypothesis that more diverse pretrain data improve a neural language model's finetuned task performance relative to random pretrain data, holding the model specifications and domain constant. A simple heuristic arising from this research project holds that an additional pretrain dataset in NLP should contain 100M words or more. Another, more tentative heuristic is that simple language models work best for selecting data inputs to more complex language models. Future research seeking to establish the merits of pretrain dataset diversity should carefully consider choices of: (a) model architecture, for example BERT or RoBERTa; (b) corpus, for example Common Crawl news or TriviaQA; (c) diversity metrics, whether those introduced by Stasaski et al. (2020), Palumbo et al. (2020), Lai et al. (2020), or others; and (d) evaluative task, for example SQuAD.

Known Project Limitations

The formulation of the "diversity hypothesis" stated above relies solely on the author's intuition and close reading of published research, not any formal proofs. All the model parameters chosen were originally preset by researchers at Google/Nvidia and Hugging Face; no additional hyperparameter search was conducted, due to time constraints.

Authorship Statement

All work reported in this article was completed by Jon Ball, a PhD candidate in Education Data Science at the Stanford Graduate School of Education. He thanks Chris Potts and Aasavari Kakne for their support.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Elvio Amparore, Alan Perotti, and Paolo Bajardi. 2021. To trust or not to trust an explanation: using leaf to evaluate local linear xai methods. *PeerJ Computer Science*, 7:e479.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2021. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1739–1746, Marseille, France. European Language Resources Association.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier detection for improved data quality and diversity in dialog systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 517–527, Minneapolis, Minnesota. Association for Computational Linguistics.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*.

Enrico Palumbo, Andrea Mezzalira, Cristina Marco, Alessandro Manzotti, and Daniele Amberti. 2020. Semantic diversity for natural language understanding evaluation in dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 44–49, Online. International Committee on Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Katherine Stasaski, Grace Hui Yang, and Marti A. Hearst. 2020. More diverse dialogue datasets via diversity-informed data collection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4958–4968, Online. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.