Exploring Effects of downsizing Parameters in ELECTRA

Kalyan Bhetwal
Department of Computer Science
Boise State University
kalyanbhetwal@u.boisestate.edu

Abstract

Transformer-based language models have been very successful in many natural language understanding (NLU) tasks. However, these models require huge datasets for training to perform well. They also require extensive hardware and computing power to complete the training. On the other hand, children are exposed to much less data than language models but do much better in NLU tasks. In this paper, we investigate using the AO-CHILDIZE dataset to train ELECTRA to explore how low we can go on model size while maintaining acceptable performance on the GLUE benchmark to that of base models. Also, we explore the effect of different hyper-parameters on the model's performance. They will help us find the best configuration of hyper-parameters without a significant hit on performance. We found that having a small hidden size and deeper network performs better while training in a small data regime. Also, we found Child-ELECTRA[1] and ELECTRA-small have comparable performance if both are pre-trained using the AO-CHILDIZE dataset while Child-Electra is 15X smaller in size.

1 Introduction

Children are fast learners. They can master their mother tongue within a few years(Leung et al., 2021). The efficiency with which children learn is quite impressive. The child-directed speech is much different from adult ones (Das et al., 1998). A child can acquire near adult-like grammatical knowledge by about the age of 6(Kemp et al., 2005) while being exposed to no more than 10-50 Million words(Hart and Risley, 1995)(Huebner et al., 2021). On the other hand, the pre-trained language Models (PLMs), which are very popular with the advent of multi-headed attention-based transformers(Vaswani et al., 2017), have transformed not only the NLP landscape, but had a significant impact on computer vision, speech processing, and many more(Lin et al., 2021), still have a significant advantage than children on the amount of data they are trained. BERT is a transformer model that employs masked-language modeling (MLM) and next sentence predication training(Devlin et al., 2018). BERT is the best-known and one of the most successful transformer models. Many models have been proposed in addition to BERT such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), ELECTRA (Clark et al., 2020), XL-Net (Yang et al., 2019) , DistilBERT (Yang et al., 2019), SPANBERT (Joshi et al., 2020), BERTSUM (Liu, 2019), and more (Rogers et al., 2020). BERT-based models are trained on a huge dataset containing billions of words. Also, they require servers that consume considerable memory and CPUs and take a significant amount of time to train the model. The BERT-Base model alone has 110 million parameters. Table 2 shows the size, training time, and amount of data used to train different models. We can establish from the table that training language models are a costly process.

ELECTRA is one of the MLMs based on BERT. Inspired by Generative Adversarial Networks (GANs), its training is based on Replaced Token Detection (RTD). Rather than corrupting the input by replacing tokens with "[MASK]" as BERT does, it corrupts the input by replacing specific input tokens with inaccurate but realistic fakes. The discriminator in the model must then decide whether tokens from the original input have been replaced or kept the same as part of the pre-training process(Clark et al., 2020).

ELECTRA achieves a new state-of-art for a single model in the SQuAD 2.0 question answering dataset (Rajpurkar et al., 2018) and outperforms

[1]Our code and pre-trained models are available at https://github.com/kalyanbhetwal/child-ELECTRA

	ELECTRA-small	Child-ELECTRA
parameters	14M	0.9 M
data size	20 MB	20 MB
max sequence	8	8
steps	60,000	60,000
hardware	v100	v100
training time	2 hours	1 hour
Accuracy	51.57	57.78

Table 1: Comparison of ELECTRA-small (Attention heads=4, Hidden layers=12 and Hidden size=256) and Child-ELECTRA (Attention heads=2, Hidden layers=6 and Hidden size=64) both trained on AO-CHILDIES dataset.

existing PLMs like RoBERTa, XLNet, and ALBERT on the GLUE score. While the large-scale T5-11b(Ni et al., 2021) model scores higher still on GLUE, ELECTRA is 1/30th its size and uses 10% of the compute power to train(Clark and Luong, 2020). Although there is a significant reduction in model size, ELECTRA-base still has 110 million parameters.

However, from children's language acquisition study in linguistics, we know that children can learn efficiently while being exposed to much lesser than PLMs like ELECTRA. So, we can say that it should be possible to construct a language model with significantly fewer parameters if it is pre-trained with a child-directed language dataset while maintaining acceptable performance. There should be a point to which we can reduce the model size without sacrificing performance in a child-centric dataset.

1.1 Aims and Hypothesis

The TLMs trained on billions of words will behave differently in a small data regime. We have two research questions in particular. First, we asked, what is the ideal parameter size of a language model for an acceptable degree of performance while being trained in a small data regime? Our second question is, how is the performance hit while downsizing the TLMs. It gives us a characteristic of the degree of loss in accuracy while downsizing the models.

1.2 Contribution

This work evaluates the training of Transformer based Language Models (TLMs) in a small data regime. Also, we studied the effect of various model parameters on model performance. There are two main contributions of this work. 1) We show that it is possible to train a language model with significantly fewer parameters and still have comparable performance to the base language model if both are trained in a small data regime. Table 1 shows that ELECTRA-small and Child-ELECTRA-small have comparable performance in GLUE task when they are both pre-trained using AO-CHIDIZE dataset, while Child-ELECTRA has 15X fewer parameters. 2) We also found having a smaller hidden layer's size and a slightly deeper network performs better.

2 Related Work

Various models and techniques have been proposed to reduce training time and model size while improving or maintaining the existing performance. DistilBERT uses knowledge distillation(KD) in the pre-training phase, reducing the size of the BERT model by 40% while retaining 97% of its language understanding capabilities and being 60% faster (Sanh et al., 2019). TinyBERT is the state-of-the-art model among models that uses KD(Jiao et al., 2019). KD is a model compression technique where a small (student) model is trained to mimic a larger (teacher) model. In this work, we did not use knowledge distillation. Instead, we used a different dataset for training. BabyBERTa trained RoBERTa model used AO-CHIDIZE dataset to mimic the input accessible to children aged 1 to 6(Huebner et al., 2021). BabyBERTa used a novel grammar test suite for evaluation. While we also used the AO-CHIDIZE dataset for pre-training, our objective is to find ideal parameters and acceptable loss in the model's performance, which is different from BabyBERTa. Ganesh et al. surveyed various BERT variants and summarized fundamental techniques being used in compressing them(Ganesh et al., 2021). They found that quantization, unstructured pruning, and structured pruning are widely used techniques. Quantization is a technique in which the number of bits used to represent a scalar parameter is reduced. A more significant part of the network, such as a channel or layer, is removed in structured pruning. However, in unstructured pruning, less meaningful connections are identified and removed. In conclusion, these works and techniques suggest that it is possible to downsize large language models while only taking a slight hit on performance; different from these works, we train ELECTRA using the AO-CHIDIZE dataset with the object of finding the best configuration of

Model	Size(millions)	Training Time	Training Data
BERT	**Base: 110** **Large: 340**	**Base: 8 x V100 x 12 days** **Large: 64 TPU Chips x 4 days** **(or 280 x V100 x 1 days)**	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.
RoBERTa	**Base: 110** **Large: 340**	**Large: 1024 x V100 x 1 day;** **4-5 times more** **than BERT.**	160 GB (16 GB BERT data + 144 GB additional)
DistiIBERT	**Base: 66**	**Base: 8 x V100 x 3.5 days;** **4 times less than** **BERT.**	160 GB (16 GB BERT + 144 GB additional) 3.3 Billion words.
XLNet	**Base:110** **Large: 340**	**Large: 512 TPU Chips x 2.5 days;** **5 times more** **than BERT.**	Base: 16 GB BERT data Large: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
ELECTRA	**Small: 14** **Base: 110** **Large: 335**	**Small: V100 x 4 days.**	12GB OpenWebTextCorpus

Table 2: Comparison of Model size, training time and training data for different PLMs (Lam, 2019)

parameters and studying the effect of those parameters on the GLUE benchmark.

3 Methods

3.1 Child-ELECTRA

We pre-trained the ELECTRA model from scratch using the AO-CHILDES dataset, and started scaling down the model to study the effect of the specific parameters on the model's performance. To fulfill our object of downsizing the ELECTRA, we had three main variables of interest (i.e. hidden layer's size, number of hidden layers, and number of attention heads). We chose these three parameters because previous work on BabyBERTa had shown that a combination of fewer layers, fewer hidden units, and fewer attention heads had grammatical understanding comparable to that of RoBERTa-base.

We named our model Child-ELECTRA since the model is derived from training ELECTRA with a child-directed language dataset. We trained the child-ELECTRA on 5 million words. The smallest model we trained has two attention heads, six hidden layers, and a hidden size of 32.

3.2 Experimental Setup

In our experiment, we changed the above three variables (i.e. hidden layer's size, number of hidden layers, and number of attention heads). Figure 1 shows the overall configuration for our study. We have three configurations in total. We pre-trained ELECTRA-small with default parameters to consider this as a baseline model. In Experiment 1, we varied hidden size while keeping the hidden layer and attention head constant. In Experiment 2, we varied hidden layers keeping hidden size and attention heads constant. We pre-trained all models except ELECTRA-small with two attention heads to avoid issues like bloated attention heads and not enough hidden layers. For pre-training, we chose a batch size of length eight to train it on a single GPU. Also, we choose the value of eight for maximum sequence length, being inspired by BabyBERTa. All of these models were pre-trained for 60,000 steps.

3.3 Training Dataset

We used Age Ordered-CHILDES (AO-CHILDES) (Huebner et al., 2021) for training the ELECTRA model from scratch. The dataset consists of approximately 5 million words that were obtained from the CHILDES database (MacWhinney, 2000). The CHILDES database was created by transcribing the various in-home recordings of children's casual speech or lab recordings of children's reading, which multiple researchers collected.

3.4 Downstream Tasks from GLUE

Dataset	#Train	#Dev	Metrics
Single-sentence Tasks			
SST-2	67k	872	Accuracy
Inference			
MNLI	393k	9.8k	Accuracy
Similarity and Paraphrase			
MRPC	3.7k	408	Accuracy/ F1

Table 3: Statistics and metrics of three dataset from GLUE benchmark.

General Language Understanding Evaluation

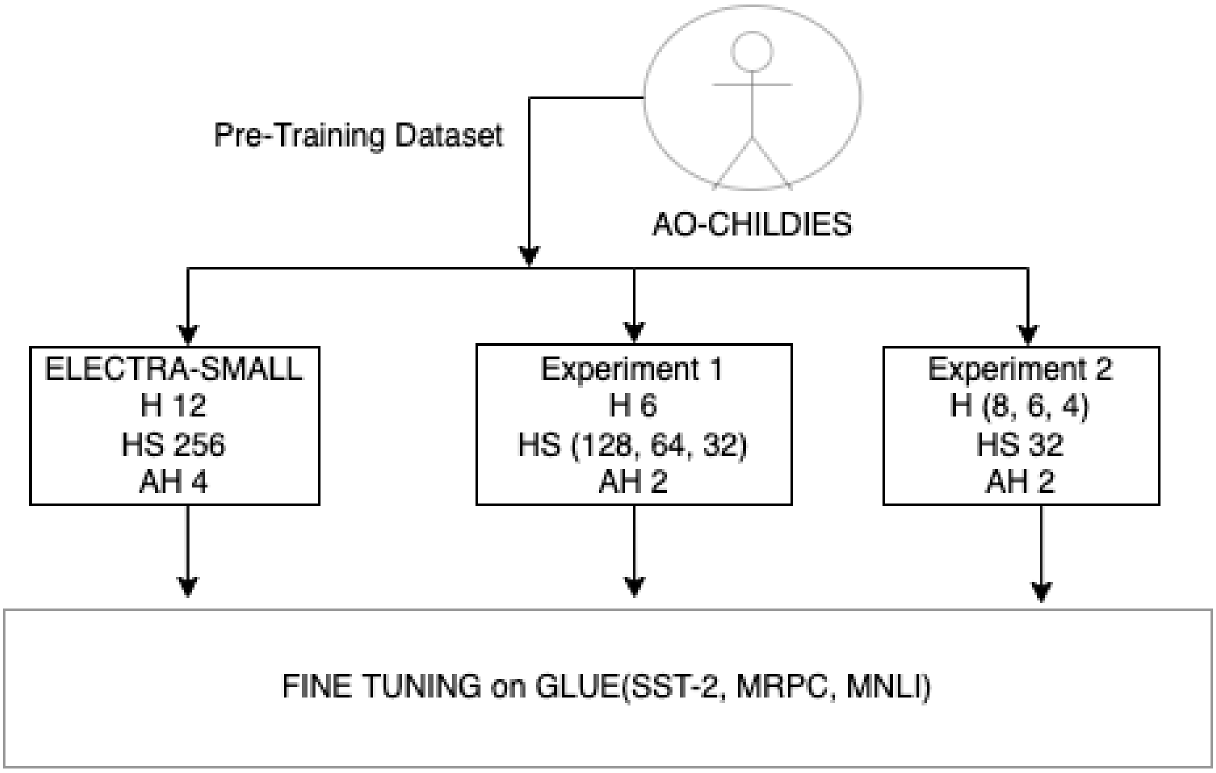

Figure 1: ELECTRA-small model is pre-trained with AO-CHILDES dataset. This model is taken as a point of reference. In Experiment 1, we changed the hidden size (128, 64 and 32) keeping hidden layers (6) and attention heads (2) constant. In Experiment 2, we changed number of hidden layers (8,6, and 4) keeping hidden size (32) and attention heads (2) constant. All of these models are benchmarked against three GLUE tasks(MNLI, MRPC and SST-2). The symbol "H" is hidden layer, "HS" is hidden size and "AH" is attention heads.

(GLUE)(Wang et al., 2018) tasks are tools for evaluating and analyzing the performance of various language models on a wide range of natural language understanding tasks. We used the GLUE benchmark because they are widely used datasets for evaluating language-oriented task performances.

For this project, we have chosen (MRPC, MNLI, and SST-2) tasks. We selected these three tasks to represent a task from three broad categories of GLUE tasks (i.e., single-sentence tasks, similarity and paraphrase tasks, and inference tasks).

Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) uses accuracy and f1 score to measure the performance on paraphrase task.

Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2018) is a natural language inference task and uses accuracy to measure the performance.

Stanford Sentiment Treebank(SST-2) (Socher et al., 2013) is single-sentence classification task and uses accuracy measure the performance.

3.5 Evaluation Method

Finally, we fine-tuned the pre-trained models on the above three GLUE downstream tasks and recorded the scores of these models. We then tabled each parameter against the GLUE benchmark and plotted graphs to see the relationship between a parameter and the corresponding GLUE score.

4 Result

4.1 Experiment 1: Changing hidden size while keeping hidden layers and attention heads fixed

Hidden Size	MRPC acc/f1	MNLI acc	SST-2 acc
128	68.38/81.3	35.46	50.92
64	68.38/81.3	48.88	77.98
32	68.38/81.3	49.82	75.11

Table 4: GLUE scores for different hidden size, fixed attention heads of 2 and hidden layers of 6

Table 3 shows the result for different hidden sizes (128, 64, and 32) while keeping the hidden

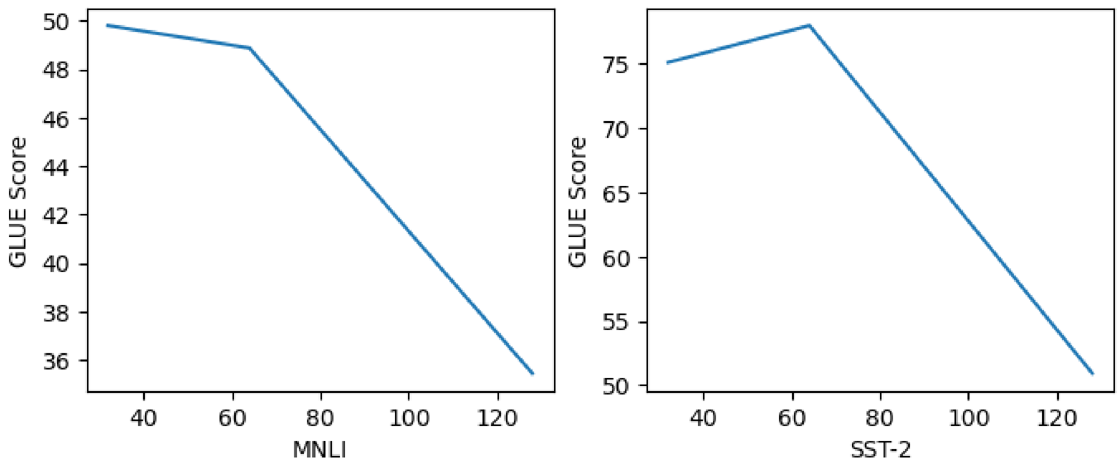

Figure 2: MNLI and SST-2 tasks score for hidden sizes of 128, 64 and 32, attention heads of 2 and hidden layers of 6

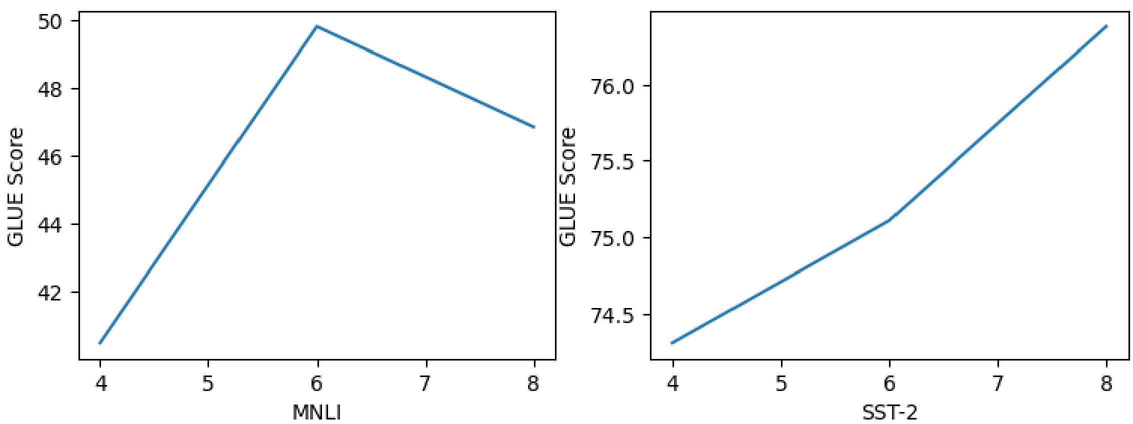

Figure 3: MNLI and SST-2 tasks score for hidden layers 8, 6 and 4, attention heads of 2 and hidden size of 32

layer's number to 6 and attention heads to 2. We can see that MRPC is 68.4 for all configurations of the hidden layer. Figure 2 shows variations of MNLI and SST-2 scores with the hidden sizes. It shows a sharp fall in performance while the hidden size is increased. The performance is comparable for the hidden size of 32 and 64 in both tasks.

4.2 Experiment 2: Changing hidden layers while keeping the hidden size and the attention heads fixed

Hidden Layers	MRPC acc	MNLI acc	SST-2 acc
8	68.38/81.3	46.85	76.38
6	68.38/81.3	49.82	75.11
4	68.38/81.3	40.49	74.31

Table 5: GLUE scores for different hidden layers, fixed attention heads of 2 and hidden size of 32

Table 4 shows the models performance for the different hidden layers(8, 6, and 4) while keeping the hidden layer's size to 32 and the attention heads to 2. We can see the MRPC value of 68.4 for all configurations of the hidden layer. Figure 4 shows the plot of MNLI and SST-2 against the different number of hidden layers. The plot of SST-2 does not vary significantly. But for MNLI, we can see the lowest score while the number of the hidden layer is small. However, we can see similar performance for the higher number of hidden layers (6 and more).

5 Discussion

5.1 Performance Characteristics

From figure 2, we can conclude that having a larger hidden size does not favor performance. It is due to the overfitting of the small dataset when there are more hidden layers. Also, from figure 3, we can

Model	MRPC	MNLI	SST-2
DistilBERT	87.5	82.2	91.3
TinyBert	87.3	84.6	93.1
ELECTRA-small	88.0	81.3	91.2
ELECTRA-large	**90.4**	**90.9**	**96.9**
ELECTRA-small-AO-CHILDIZE	68.4	35.4	50.91
Child-ELECTRA-best	68.4	49.82	75.12

Table 6: Comparison of various PLMs. ELECTRA-small-AO-CHILDIZE (Attention heads=4, Hidden Size=256, and Hidden Layer=12) is ELECTRA-small trained on AO-CHIDIZE and Child-ELECTRA-best (Attention heads=2, Hidden Size=32, and Hidden Layer=6) is best performing model for all configuration in our study

conclude that having a deeper network favors the language model. Decreasing the value below a certain level causes underfitting, and having excessive hidden layers causes overfitting. We should consider both of these parameters in unison. We found that hidden size of 32, hidden layers of 8, and attention heads of 2 perform best for our configuration. Also, we found that having fewer attention heads favors performance due to our use of smaller input sequences. We saw that MPRC was 68.4 and constant for all the configurations. While fine-tuning, we found that all of the cases in test cases were categorized as positive. The MRPC dataset has 68.4 positive cases. Hence, the final accuracy was always 68.4.

5.2 Comparison to Existing Models

Table 6 shows a comparison of our work with existing language models. Our work did not perform well on GLUE benchmarks compared to ELECTRA-large and ELECTRA-small models. It was expected since training a larger model on billions of words is different from training a smaller model with a smaller dataset. However, ELECTRA-small-AO-CHILDIZE (Attention heads=4, Hidden Size=256 and Hidden Layer=12) trained in AO-CHILDIZE has comparable performance to Child-ELECTRA-best (Attention heads=2, Hidden Size=32 and Hidden Layer=6). From this observation, we can say that a significantly downsized model can have consistent or comparable performance to the base model if both are trained in a small data regime.

5.3 Limitations and Future Directions

The main limitation of our work stems from the fact that although we trained ELECTRA for a different combination of attention head, hidden size, and hidden layers, we only explored limited combinations. Also, we did not explore other hyperparameters like intermediate size and other influential parameters r that could have helped us improve accuracy. Although we used a Child language dataset, we did not build the model to represent the Child language acquisition process. Using a specific dataset might add some value, but building-specific models that mimic the actual learning process of children could be a deal-breaker and might eventually reduce the model size while giving an acceptable performance. We used GLUE for evaluation, but it would make more sense if there were a child-centric evaluation dataset.

In ELECTRA, the generator has to be scaled as per the changes in the discriminator. Although we changed the parameters in the discriminator, the generator was left as it is. The recommended generator size should be roughly a quarter to half of the discriminator's size for practical training. If the generator size is more than half of the discriminator's size, the generator will be too good, and the adversarial game will collapse. It should be considered in future work..

6 Conclusion

The main contribution of this work is exploring how PLMs perform in a small data regime and observing how specific parameters influence the overall performance of PLMs. We found that having a fewer hidden size and deeper networks helps language models to perform better in a small data regime. Also, We found that ELECTRA-small has comparable performance with CHILD-ELECTRA if we train both of them with the AO-CHILDIZE dataset, while CHILD-ELECTRA has 15X fewer parameters than ELECTRA-small. Thus, we can conclude that having fewer parameters still has comparable performance to the larger language models in a small data regime.

7 Acknowledgement

We are thankful to the anonymous reviewers for their valuable feedback, which helped to improve the quality of this work. Also, we would like to thank Boise State Super Computing facility for providing computing cluster.

References

K Clark and T Luong. 2020. More efficient nlp model pre-training with electra. *Preuzeto s from https://ai. googleblog. com/2020/03/more-efficient-nlp-model-pre-training. html [4. srpnja 2021.]*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Subrata Das, Don Nix, and Michael Picheny. 1998. Improvements in children's speech recognition performance. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 1, pages 433–436. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. 2021. Compressing large-scale transformer-based models: A case study on bert. *Transactions of the Association for Computational Linguistics*, 9:1061–1080.

Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Philip A Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. Babyberta: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Nenagh Kemp, Elena Lieven, and Michael Tomasello. 2005. Young children's knowledge of the" determiner" and" adjective" categories.

Eric Lam. 2019. What happens after bertnbsp;? summarize those ideas behind.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Ashley Leung, Alexandra Tunkel, and Daniel Yurovsky. 2021. Parents fine-tune their speech to children's vocabulary knowledge. *Psychological Science*, 32(7):975–984.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. A survey of transformers. *arXiv preprint arXiv:2106.04554*.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brian MacWhinney. 2000. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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)*, pages 1112–1122. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.