Sparse*BERT: Sparse Models are Robust

Daniel Campos1,2, Alexandre Marques2, Tuan Nguyen2, Mark Kurtz2, and ChengXiang Zhai1

1Department of Computer Science, University of Illinois Urbana-Champaign
2Neural Magic Inc.

Abstract

Large Language Models have become the core architecture upon which most modern natural language processing (NLP) systems build. These models can consistently deliver impressive accuracy and robustness across tasks and domains, but their high computational overhead can make inference difficult and expensive. To make the usage of these models less costly recent work has explored leveraging structured and unstructured pruning, quantization, and distillation as ways to improve inference speed and decrease size. This paper studies how models pruned using Gradual Unstructured Magnitude Pruning can transfer between domains and tasks. Our experimentation shows that models that are pruned during pretraining using general domain masked language models can transfer to novel domains and tasks without extensive hyperparameter exploration or specialized approaches. We demonstrate that our general sparse model Sparse*BERT can become SparseBioBERT simply by pretraining the compressed architecture on unstructured biomedical text. Moreover, we show that SparseBioBERT can match the quality of BioBERT with only 10% of the parameters.

1 Introduction

Foundational Models (Bommasani et al., 2021) based on the Transformer architecture (Vaswani et al., 2017) have quickly become the most common building block in the modern language understanding stack. While these models are able to provide robust language representations which can be leveraged by to provide impressive accuracy on tasks like question answering, text classification, and token classification. These Large Language Models (LLMs) have shown to be robust to shifts in domain and domain specific models like BioBERT (Lee et al., 2020), LEGALBERT (Chalkidis et al., 2020), and SciBERT (Beltagy et al., 2019) have become a popular strategy for improving performance further. While accurate and robust, LLM are not without drawbacks. These Large Language Models (LLMs) commonly have hundreds of millions or billions of parameters which commonly require large specialized computer clusters to run inference at scale. Several approaches have been successfully used to improve performance of these LLMs, such as approximating attention (Peng et al., 2021), removing portions of the models (Sridhar and Sarah, 2020) and reducing the precision of activation and weight values.

Recent work (Zafrir et al., 2021) (Kurtič et al., 2022) has shown that the application of unstructured and semi-structured (block) pruning mechanisms on LLMs can significantly compress models with little to no loss in accuracy. While these approaches are successful and applicable during model general pretraining and task-specific fine-tuning, prior work has not studied how pruned models transfer to novel domains nor the impact of pretraining stage pruning on transfer accuracy. In our work, we focus on studying how generalizable
pruned LLMs are by answering the following research questions:

- Can models sparsified using general domain text transfer to new domains without further optimization?
- Given a novel textual domain and transfer tasks, where can Gradual Magnitude Pruning best be applied?
- Can domain-specific pruned LLMs provide performance equal to the uncompressed models?

We explore these questions by focusing on transferring pruned and unpruned LLM to the biomedical domain and evaluating the accuracy of said models on downstream tasks like Entity Extraction (EE) and Relation Extraction (RE), and Question Answering (QA). Our experiments demonstrate that pruned models generalize well and are robust to domain transfer and variation in target task, dataset size, and dataset difficulty. In summary, our contributions are as follows:

- We extensively evaluate how Gradual Magnitude Pruning performs when applied to LLMs transferred to novel domains and tasks.
- We demonstrate that LLMs pruned on the general domain language can transfer to novel domains without extensive hyperparameter tuning and can produce equivalent or better accuracy when compared to dense LLMs.
- We introduce a sparse model, Sparse*BERT, and its domain adaptation adapted for the Medical/Bio NLP domain called SparseBioBERT. This model matches the accuracy of the BioBERT with 10% of its parameters pruned runs inference 4 times smaller with no drop in accuracy.
- We introduce a sparse model, Sparse*BERT, and its domain adaptation adapted for the Medical/Bio NLP domain called SparseBioBERT. This model matches the accuracy of the BioBERT with 10% of its parameters pruned runs inference 4 times smaller with no drop in accuracy.

2 Background and Related Work

Large Language Models such as language representation and generation models commonly use multiple layers of transformer encoders or decoders. Each transformer layer usually contains some form of multi-head attention (MHA) and fully-connected feed-forward networks (FFN). The MHA is made up of multiple self-attention heads (Vaswani et al., 2017) each which has 3 sub-components: queries (Q), keys (K), and values (V). Equation 1 shows the expression used to compute the attention of each head, where \( d \) is the dimensionality of \( K \). The output of each attention head is concatenated together and fed into the FFN.

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK}{\sqrt{d}} \right) V
\]  

Attention, while simple, has proven to be incredibly robust as it allows models to scale to hundreds of layers, hundreds of attention heads (Brown et al., 2020), and seemingly most modalities (Chen et al., 2021) (Arnab et al., 2021). Despite its ability to generalize, attention-based models are also brittle as removing less than 0.0001% of parameters can cause complete model collapse (Kovaleva et al., 2021a).

Unstructured pruning compresses a model by removing individual weights from a network by setting them to zero. Pruning methods commonly remove weights based on their saliency to the network and, to avoid model collapse, usually do so gradually while fine-tuning the remainder of the weights. Since it is difficult to quantify the true saliency of weight concerning a network, zeroth, first, and second-order estimation methods exist to approximate saliency.

Zero-order methods use weight magnitudes as a proxy, i.e., remove the smallest weights without evaluating the impact of their removal on model accuracy. These approaches are prevalent for Convolution Neural Networks (Han et al., 2015) and have recently been successful for LLM (Zafrir et al., 2021). First-order methods like Movement Pruning (Sanh et al., 2020) use a gradient-based approximation to remove the weights that are moving toward zero. Second-order methods like OBS (Singh and Alistarh, 2020) estimate the impact of individual weight removal via approximations of second-order derivatives and use it as a proxy for saliency.

The use of unstructured and semi-structured pruning has proved to be a convenient way of compressing LLM for efficient inference and decreased model size. For example, a BERT-Base-Uncased model which has had 90% of its parameters pruned runs inference \( \sim 4.5 \) times faster and is 2.75 times smaller with no drop in accuracy (Kurtić et al., 2022). If this compression leverages additional methods like quantization and layer dropping, inference speed can improve by 28.75 times, and model size can drop by \( \sim 19 \) times.

While many successful compression approaches exist, Transformer models are fragile (Kovaleva et al., 2021b), as minor perturbations can lead to model collapse.

3 Sparse*BERT: General Sparse Models Can Adapt to Novel Domains

While existing pruning research has found it possible to prune models heavily without loss in accuracy, most approaches have focused on the compression of individual tasks or textual domains. These specialized models match or exceed the accuracy of the dense model but commonly require vast amounts of hyperparameter tuning and task-specific optimization to achieve this result. Models like DistillBERT (Sanh et al., 2019), and TinyBERT (Jiao et al., 2020) are some of the most popular
compressed LLMs because they provide compression without any additional know-how or optimization. To garner widespread adoption of sparse models, the creation of pruned models which can adapt to new tasks and domains without additional tuning is paramount. Our methods focus on producing Sparse*BERT, a pruned LLM that adapts to novel domains without extensive fine-tuning or task-specific pruning.

We can formulate the Sparse*BERT model as \( \theta^* \), which can approximate the accuracy of the dense model \( \theta \) and does not suffer model collapse when transferred to novel domains. The architecture of Sparse*BERT matches BERT, but a portion of its weights are pruned and masked to avoid future updates. To ensure that our sparse model can approximate the accuracy of the dense model, we leverage Knowledge Distillation by comparing the relative entropy using the Kullback–Leibler divergence between the outputs of the dense and sparse networks.

Following the success of Zafrir et al. (Zafrir et al., 2021) we leverage Gradual Magnitude Pruning (GMP) as shown in algorithm 1. To ensure that we can maximize the inference speedups, we prune each set of components in the network independently. The structure in the model graph groups these components, so the individual feed-forward layers, the queries, or keys, but not individual self-attention heads.

Algorithm 1 Uniform Gradual Magnitude Pruning

Input: \( \theta_0 \), a pretrained neural network, \( \theta_t \), a dense pretrained neural network (distillation teacher), \( D \) a training dataset, \( N \) number of pruning steps, \( \sigma \) weights to prune at each pruning step

Output: a pruned neural network

for \( x \) in 1 to \( N \) do

\( \theta^* \leftarrow \theta_0 \)

for component in \( \theta^* \) do

\( w \leftarrow \text{sort}() \)

\( \theta^* \leftarrow \text{prune}(\theta^*, w, \sigma) \)

end for

\( \theta^* \leftarrow \text{train}(\theta^*, \theta_t, D) \)

end for

return \( \theta^* \)

4 Experiments

To assess how well-pruned models can transfer to novel domains and determine the optimal stage of pruning (pretraining, domain transfer, or fine-tuning), we evaluate the accuracy on 10 Biomedical biomedical datasets. Our experiments fix the training parameters for all the transfer tasks and vary the stage used for pruning (no pruning, pretraining, domain transfer, fine-tuning) and domain-specific pretraining.

4.1 Datasets

Pretraining Datasets. To understand how the stage of pruning impacts model accuracy, we train models both pruned and dense models on the Medline/PubMed corpus and the combination of English Wikipedia (Foundation, 2021) and The Book Corpus (Zhu et al., 2015) datasets. The combination of Wikipedia and Book Corpus dataset creates a common domain language dataset featuring 3.3 billion words which have become the backbone for experimentation for general domain masked language modeling.

The MEDLINE/PubMed corpus is a publicly available\(^1\) text corpus made up of journal abstracts and documents of biomedical literature from around the world. The corpus is updated daily by the United States National Institute of Health and it has been the primary resource used to train BioMedicaal LLMs like BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2022). For our experiments, we extracted our corpus on January 2022, and filter and prepare the dataset for masked language modeling using the BioElectras (Kanakarajan et al., 2021) scripts\(^2\). This formatted PubMed corpus has 34,246,348 abstracts and 4.5 billion words (Kanakarajan et al., 2021).

Finetuning Datasets. We finetune pretrained models on 10 established BioMedical NLP datasets, encompassing 3 separate task types: Entity Recogniton (ER), Relation Extraction (RE), and Question answering (QA). For ER we use the BioCreative II Gene Mention Recognition (BC2GM), (Smith et al., 2008), BC5CDR Drug/Chemical (BC5-Chem), BC5CDR Disease (BC5-Disease) (Li et al., 2016), JNLPBA (Collier and Kim, 2004), and NCBI Disease (Dogan et al., 2014) datasets. For RE we use ChemProt (Taboureau et al., 2011), Drug Disease Interaction (DDI) (Herrero-Zazo et al., 2013), and Gene-Disease Associations (GAD) (Becker et al., 2004) datasets. For QA we leverage BioASQ task 7B (Baker et al., 2016) and PubMedQA (Jin et al., 2019). In addition, we perform an analysis of the impact of the size of the finetuning dataset on the optimal stage for pruning using the non biomedical QA SQUAD (Rajpurkar et al., 2016) dataset. Details on dataset size, evaluation metric and domain can be found in table 1.

4.1.1 Models and Experimental Setup

Our experiments focus on the popular BERT-base-uncased language model (Devlin et al., 2019), which is an LLM composed of 12 transformer encoder layers and has 110M parameters. Following previous work, we do not prune the embedding layers of the network or any task-specific heads and focus on the ~85 million

\(^1\)https://www.nlm.nih.gov/databases/download/pubmed_medline.html

\(^2\)https://github.com/kamalkraj/BioNLP-Corpus
Table 1: In order to understand how generalizable sparse models are we evaluate on a wide set of tasks that vary in difficult, size, and desired output

| Dataset      | Domain | Task Type                  | Training Size | Testing Size | Validation Size | Evaluation Metric |
|--------------|--------|----------------------------|---------------|--------------|-----------------|-------------------|
| BC5-Chem     | Medical| Entity Recognition         | 5203          | 5347         | 5385            | F1                |
| BC5-disease  | Medical| Entity Recognition         | 4182          | 4244         | 4424            | F1                |
| NCBI-disease | Medical| Entity Recognition         | 5134          | 787          | 960             | F1                |
| I2BMED       | Medical| Entity Recognition         | 15197         | 3061         | 6325            | F1                |
| ChemProt     | Medical| Entity Recognition         | 46750         | 4551         | 8662            | F1                |
| DDI          | Medical| Relation                   | 18035         | 11268        | 15745           | F1                |
| GAD          | Medical| Relation                   | 25296         | 2496         | 5716            | F1                |
| PubMedQA     | Medical| Question Answering         | 4263          | 535          | 534             | Accuracy          |
| BioASQ       | Medical| Question Answering         | 760           | 75           | 140             | Accuracy          |
| SQUAD        | General| Question Answering         | 87599         | 10570        | N/A             | F1                |

Table 2: Overall results on the impact of task and dataset of model pruning. Models trained for the general domain and pruned on the general domain can transfer at equal or better accuracy. Question Answering is the notable outlier as its small dataset size benefits from the sparse models as their pruned architecture prevents overfitting on small datasets.

| Model                  | Pruning Stage | EE | RE | QA | Overall |
|------------------------|---------------|----|----|----|---------|
| BioBERT-Base-Uncased   | None          | 85.96 | 87.56 | 68.33 | 80.62   |
|                        | Fine pruning  | 63.53 | 75.14 | 54.00 | 66.34   |
|                        | Pretraining   | 82.50 | 86.60 | 65.71 | 78.27   |
|                        | Pretraining General language | 86.36 | 88.57 | 66.33 | 80.42   |

| Model                  | Pruning Stage | EE | RE | QA | Overall |
|------------------------|---------------|----|----|----|---------|
| BERT-Base-Uncased      | None          | 84.44 | 86.86 | 61.97 | 77.76   |
|                        | Fine pruning  | 76.12 | 85.36 | 65.39 | 75.28   |
|                        | Pretraining   | 80.84 | 85.54 | 68.44 | 78.27   |

4.2 Model Pretraining

Pretraining refers to the stage in which the model is trained on an unsupervised NLP dataset using a masked language modeling (MLM) approach (Devlin et al., 2019). Pretrained models are fine-tuned on labeled task-specific datasets to optimize for task-specific accuracy. For our experiments, we use existing dense pretrained models for BERT-base-uncased (Devlin et al., 2019) and PubmedBERT (Gu et al., 2022) and prune them using gradual magnitude pruning based on the corresponding dataset and MLM approach. We keep the setup for pruning on the pretraining stage as consistent as possible. Namely, we train for three epochs on 4 A100 GPUS using a batch size of 256, the sequence length of 512, and, following early experiments and findings from Kurtic et al. and Zafrir et al., we cycle the learning rate during pretraining and found cycling twice per epoch from 5e-4 to 0 to be most effective. We apply gradual magnitude pruning over the first two epochs, starting from an initial sparsity of 50% and gradually pruning to a final sparsity of 90% pruning 100 times an epoch. After model pruning, we continue to train for one additional epoch to ensure that the sparse model is converged. For our dense models, we follow the same procedure but without the use of pruning.

Based on early experiments, we find knowledge distillation beneficial. For all of our experiments in pretraining, we leverage well-trained dense teachers using a hardness of 0.5 and a temperature of 2. When pruning weights, their values are fixed to 0 and are masked to avoid future updates. Effectively this means that our experiments evaluate the discovery of the most optimal sub-architecture.

4.3 Biomedical Finetuning

To isolate the effects of task-specific hyperparameter tuning and pruning stages, we fix the training procedure across fine-tuning tasks. Specifically, we train each model for ten epochs on a V100 GPU using a batch size of 16, a learning rate that linearly decays from 5e-5, and replicates using five random seeds for larger tasks and ten random seeds for smaller tasks.

We use the same setup both for fine-tuning already pruned models and when applying gradual magnitude pruning during fine-tuning (pruning on the fine-tuning stage). For pruned models, we preserve the sparsity patterns. When pruning models during fine-tuning, we fine-tune the dense model for two epochs, prune over the preceding six epochs, and stabilize the pruned network for two epochs. In our early experiments and matching prior findings (Zafrir et al., 2021), we find that when pruning models on transfer tasks, accuracy is...
best when the learning rate cycles. Cycling only occurs when pruning during fine-tuning, and the learning rate cycles at epochs 2 (start of pruning) and 8 (end of pruning). Unlike previous work, we do not find a significant effect in accuracy improvement by leveraging knowledge distillation on the fine-tuning task. As a result, we do not use knowledge distillation during fine-tuning.

4.4 Experimental Results

When we compare the impact of the pruning task on downstream model performance, we can see that there is a huge variation in performance. When we evaluate the impact of pruning on individual tasks, pruning in fine-tuning stage) as shown in table 2, we can see that pruning is quite sensitive to the dataset task. Looking at the large datasets like JNLPBA in table 3, there is nearly no distinction in pruning during pretraining or fine-tuning. On the other hand, small datasets like NCBI and GAD see a large accuracy loss from models pruned during fine-tuning.

Our results also indicate that it is optimal to prune on general domain language modeling and transfer to novel tasks and domains for optimal performance. Regardless of their domain expertise, BERT and BioBERT both see huge losses inaccuracies when pruned on the downstream tasks, and these same losses are not found in the model pruned during pretraining. Surprisingly, the model pruned on general domain language modeling outperforms when pruned on the biomedical domain-specific language modeling. This gap is nearly 4 points on entity extraction and 2 points overall, almost more significant than the gap between the BERT and BioBERT.

We believe that these results provide evidence that models pruned during general domain language modeling can remove large portions of the model without affecting the ability to transfer to novel domains or tasks. Unlike pruning on specific domains and tasks, general domain pruning can preserve accuracy invariant to task and domain.

Unexpectedly, when evaluating biomedical QA, we improve accuracy with pruned models when pruned but only on a regular BERT model pruned downstream or the general domain Sparse*BERT.

We attribute this to the regularizing effect that pruning can have, and it likely helps in overfitting on small datasets, PubMedQA and BioASQ. Tasks. Finding that models pruned during outperform all others, we believe the regularization provided by pruning can prevent the overfitting on these small datasets.

4.5 Impact of Training Data Size

Noting that there is a significant variation in dataset size in the biomedical NLP tasks, we leveraged a dataset well studied in pruning, SQUAD, and performed variations to the training data size. Starting with the initial training dataset size, 88,000 items, we decreased the size to 75%, 50%, 25%, 10%, 5%, 2%, 1% and evaluated the impact. We compared the dense BERT, Sparse*BERT, and pruning BERT during fine-tuning. The sparse models each have 90% unstructured sparsity on the encoder portion of the LLM. Each experiment was performed with five random seeds, using a batch size of 12, and trained for 30 epochs with a learning rate decaying linearly from 5e-5 to 0. For the model pruned during fine-tuning, we cycle the learning rate at the beginning of pruning (2nd epoch) and the end of pruning (20th epoch). We evaluate model performance on the F1 score on the unaltered dev portion of the SQUAD dataset to avoid losses in evaluation sensitivity.

As shown in figure 2 and table 4 results in the appendix, models which are pruned during finetuning are not robust to variations in data size. Model performance decays slowly from 85 to 80 until the training data is decreased by 75%, but when it becomes smaller than that, it quickly becomes nearly unusable. The same cannot be said about the dense or the Sparse*BERT model as they see virtually identical losses in quality from

5 Conclusion

In this work, we have introduced Sparse*BERT, a pruned LLM which builds on successful pruning algorithms research and demonstrates its ability to transfer to novel domains and tasks without additional hyperparameter search. Our experiment demonstrates how well Sparse*BERT can transfer to the biomedical domain
Table 3: Performance on Complete set of tasks. Except for question answering tasks and NCBI-Disease, the SparseBioBERT outperforms all other models, including BioBERT, indicating that sparse architectures can be transferred to novel domains and use cases without additional optimization.

| Training dataset size | N/A | 4182 | 5203 | 5134 | 15197 | 46750 | 18035 | 25296 | 4261 | 450 | 670 |
|-----------------------|-----|------|------|------|-------|-------|-------|-------|------|-----|-----|
| BERT-Base-Uncased     | None | 80.60 | 91.23 | 85.66 | 81.97 | 81.56 | 88.19 | 94.35 | 78.05 | 47.46 | 76.65 |
| Fine tuning           | 69.87 | 81.72 | 75.57 | 74.27 | 79.57 | 85.41 | 92.72 | 74.88 | 52.67 | 78.11 |
| Pretraining           | 75.35 | 87.83 | 81.75 | 77.12 | 81.24 | 86.13 | 92.73 | 77.77 | 50.00 | 82.89 |
| BioBERT-Base-Uncased  | None | 83.195 | 93.63 | 83.46 | 86.05 | 84.10 | 90.66 | 95.01 | 77.02 | 54.00 | 82.67 |
| Fine tuning           | 66.34 | 66.60 | 52.69 | 59.71 | 62.15 | 80.28 | 87.60 | 57.55 | 34.88 | 60.00 |
| Pretraining(General)  | 83.32 | 93.81 | 84.15 | 87.04 | 81.84 | 90.71 | 95.02 | 79.90 | 51.50 | 81.17 |
| Pretraining(Medical)  | 80.56 | 92.17 | 78.84 | 81.11 | 81.81 | 88.02 | 93.99 | 77.77 | 49.14 | 82.28 |

Table 4: Model accuracy as measured by F1 on dev portion of SQUAD compared to model type and the impact of training data size. Sparse models are not as sample efficient as their dense counterparts, but Sparse*BERT performance matches the dense model much more than the model pruned downstream.

| Training Data Portion | 100 | 75 | 50 | 25 | 10 | 5 | 2 | 1 |
|-----------------------|-----|----|----|----|----|---|---|---|
| BERT-BASE-Uncased     | 89.30 | 89.65 | 89.33 | 88.17 | 86.04 | 84.06 | 80.00 | 72.58 |
| Sparse*BERT           | 88.50 | 88.91 | 88.21 | 86.48 | 83.45 | 79.99 | 72.83 | 60.91 |
| Finetune Pruning      | 85.85 | 86.17 | 85.11 | 81.5  | 75.24 | 73.13 | 69.88 | 39.35 |

6 Future work

We seek to expand Sparse*BERT to novel domains such as legal, financial, and medical in future work. Furthermore, we like to continue our work on more complex models studying how sparsity impacts multilingual and multi-task models. In particular, we seek to understand how structured and unstructured approaches in compression relate to the curse of multilingualism.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schimid. 2021. Vivit: A video vision transformer. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6816–6826.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högborg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics, 32:3432–40.

Kevin G. Becker, Kathleen C. Barnes, Tiffani J. Bright, and S Alex Wang. 2004. The genetic association database. Nature Genetics, 36:431–432.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In EMNLP.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannie Bohg, Antoine Bosselot, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. 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 F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Minna Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lia Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suivir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanay, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Niforoshan, J. F. Nyarko, Giray O gut, Laurel Orr. Isabel Papadimitriou, Joon Sung Park, Chris Pech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R’e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy
**Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258.**

**Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. The optimal bert surgeon: Scalable and accurate second-order pruning for large language models. *ArXiv*, abs/2203.07259.**

**Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The puppets straight out of law school. *ArXiv*, abs/2010.02559.**

**Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 347–356.**

**Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at jnlpba. In *NLPBA/BioNLP*.**

**J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.**

**Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.**

**Wikimedia Foundation. 2021. *Wikimedia downloads*.**

**Yuxian Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.**

**Song Han, Huizi Mao, and William J. Dally. 2015. A deep neural network compression pipeline: Pruning, quantization, huffman encoding. *ArXiv*.**

**María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of biomedical informatics*, 46 5:914–20.**

**Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. *ArXiv*, abs/1909.10351.**

**Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP*.**

**Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaiyannan Sankarasubbu. 2021. Bioelectra: pre-trained biomedical text encoder using discriminators. In *BIOSLP*.**

**Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021a. Bert busters: Outlier dimensions that disrupt transformers. In *FINDINGS*.**

**Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshicky. 2021b. BERT busters: Outlier layernorm dimensions that disrupt BERT. *CoRR*, abs/2105.06990.**

**Eldar Kurtić, Daniel Fernando Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Ben Fineran, Michael Goín, and Dan Alistarh. 2022. The optimal bert surgeon: Scalable and accurate second-order pruning for large language models. *ArXiv*, abs/2203.07259.**

**Jinhuyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.**

**Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Scialy, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.**

**Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. 2021. Random feature attention. *ArXiv*, abs/2103.02143.**

**Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.**

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

**Victor Sanh, Thomas Wolf, and Alexander M. Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. *ArXiv*, abs/2005.07683.**
Sidak Pal Singh and Dan Alistarh. 2020. Woodfisher: Efficient second-order approximation for neural network compression. *Advances in Neural Information Processing Systems*, 33.

Larry L. Smith, Lorraine K. Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, C. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence E. Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter W. Adriaans, Christian Blaschke, Rafael Torres, Mariana L. Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9:S2–S2.

Sharath Nittur Sridhar and Anthony Sarah. 2020. Undivided attention: Are intermediate layers necessary for bert? *ArXiv*, abs/2012.11881.

Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S. Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skot Jensen, Søren Brunak, and Tudor I. Oprea. 2011. Chemprot: a disease chemical biology database. *Nucleic Acids Research*, 39:D367–D372.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Ofir Zafrir, Ariel Larey, Guy Boudoukh, Haihao Shen, and Moshe Wasserblat. 2021. Prune once for all: Sparse pre-trained language models. *ArXiv*, abs/2111.05754.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.