Pruning Large Language Models

Amrita Nair

March 25, 2023

1 Introduction

In recent years, the general trend has been to increase the number of parameters in large scale models to improve the performance. GPT-1, released in 2018 had 117 million parameters Radford et al. [2018], GPT-2 released in 2019 had 1.5 billion parameters Radford et al. [2019] and GPT-3 released in 2020 has 175 billion parameters Brown et al. [2020]. Better pre-training models lead to higher accuracy on a number of downstream tasks. However, higher number of parameters translates to longer compute times (Figure 1) and much higher GPU usage. This can be fixed by **pruning** the model to obtain a sub-network that contains the "essence" of the model.

Further, research in the direction of the Lottery Ticket Hypothesis(section 2) show that "winning tickets" can be universal; tickets that are generated on a particular dataset/optimixer/architecture combination can transfer to other such combinations to provide similar/comparable results. This suggestion is especially relevant in the case of Large Language Models(LLM's) like BERT or GPT which are fine-tuned on a variety of downstream tasks.

This report first provides a short description of the Lottery Ticket Hypothesis following which there is a discussion regarding the universality of lottery tickets especially for models that deal with natural image datasets. This report then focuses on Gordon et al. [2020] who describe the compressing of BERT, in the context of the Lottery Ticket Hypothesis, and state that *30-40%* of the parameters in the BERT model can be discarded based on empirical evidence that is obtained after fine-tuning on a variety of tasks.

Finally, attention is turned towards pruning a BERT(fine-tuned on a reading comprehension task) model that is then fine-tuned on other smaller datasets that deal with reading comprehension. This is a domain-adaptation task and the results show that the model introduced by Zhu et al. [2021] performs better than the full model on most of the target domains.

2 The Lottery Ticket Hypothesis

Frankle and Carbin [2018] suggested the Lottery Ticket Hypothesis(referred to as LTH in the rest of this report). The "lottery ticket" phrase refers to the sub-networks here that have won the "lottery" in terms of having connections that have initial weights that make training effective. The claim posed by Frankle and Carbin [2018] goes as follows;

The Lottery Ticket Hypothesis. *A randomly-initialized, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.*

Formally defined, consider a feed-forward neural network $f(x;\theta)$ with maximum validation loss l, at iteration j, and accuracy a. Also, consider mask $m \in 0, 1^{|\theta|}$ which modifies the initialization to give $f(x; m \odot \theta)$. The hypothesis predicts that, f reaches minimum validation loss $l^{'}$, at iteration $j^{'} \leq j$ with accuracy $a^{'} \geq a$

Frankle and Carbin [2018] provide a basic algorithm that demonstrates this hypothesis:

1. Initialize a feed-forward neural network with random parameters

2. train the network to convergence

3. prune $x\%$ of the parameters

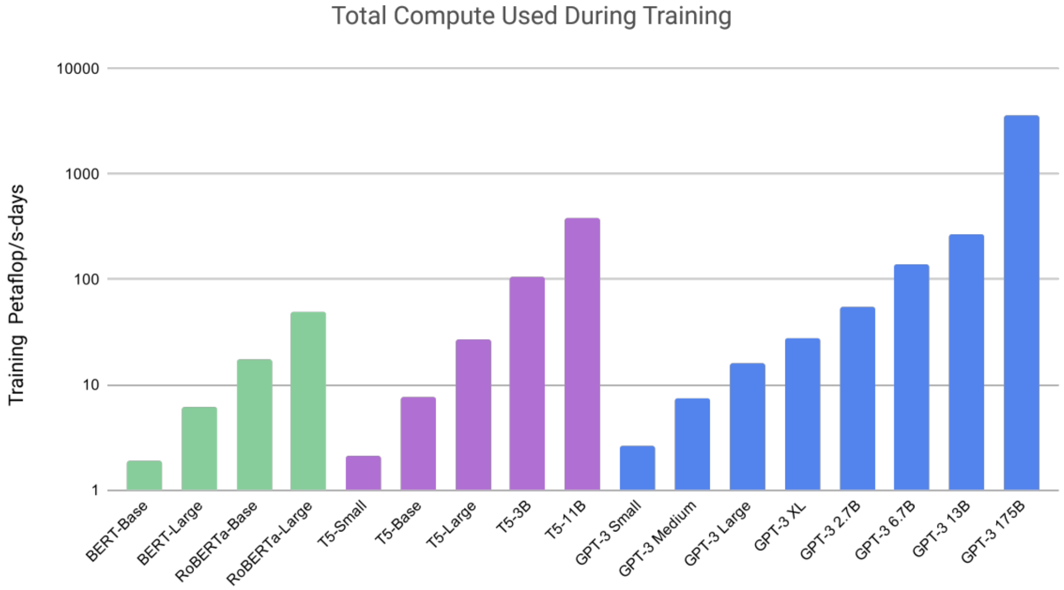


Figure 1: Brown et al. [2020]

4. **reset** the values of the remaining parameters to their original random values which would create the winning ticket, $f(x; m \odot \theta)$.

The method above is *one shot*, here the weights are pruned just once at the end of training. However, as stated in Lange [2020], *weight magnitude is often only a noisy proxy for weight importance*. One shot pruning can leave the network vulnerable to this noise. Hence, Iterative Magnitude Pruning(train-prune-train) is, in general, preferred since this protects the network from losing important weights. It involves pruning a certain percentage in every iteration. Through most of the research around LTH, different pruning paradigms are used which influence the results.

3 Universal Lottery tickets

The formulation of LTH brought to the forefront two major concerns. First, this indicates that most initialization schemes that may have been heuristically chosen are sub-optimal. Secondly, it suggests that over-parameterization is not necessary, rather, there needs to be an effort to find a good intialization scheme. In fact, over-parameterization is necessary to reach an appropriately parameterized network. This implies that training and performing inference in networks that are 1-2 orders of magnitude larger than necessary wastes large amounts of computation.

These points are put forward by Morcos et al. [2019] who conclude that a more ”principled” initialization scheme must be formulated. However, repeated pruning in computationally expensive, especially for huge models like LLM's. Further, it is not clear what properties of winning tickets make them ”winning”. Is it a special combination of architecture+dataset+optimization scheme, or is it because winning tickets contain inductive biases which improve training?

It is important to know which of these two scenarios leads to the utility of winning ticket. If the unique combination of architecture, dataset and optimization scheme give rise to winning tickets, then it would be necessary to generate winning tickets for each combination which would be very expensive, both computationally and effort-wise. On the other hand, if winning tickets inculcate generic inductive biases, there is hope that the same winning ticket can generalize well over different combinations of datasets and architectures. This would translate to saving in costs and would allow the winning tickets to be reused. Most importantly, these winning tickets can be further parameterized to generate dataset specific tickets which is a far cheaper alternative compared to over-parameterizing a huge network.

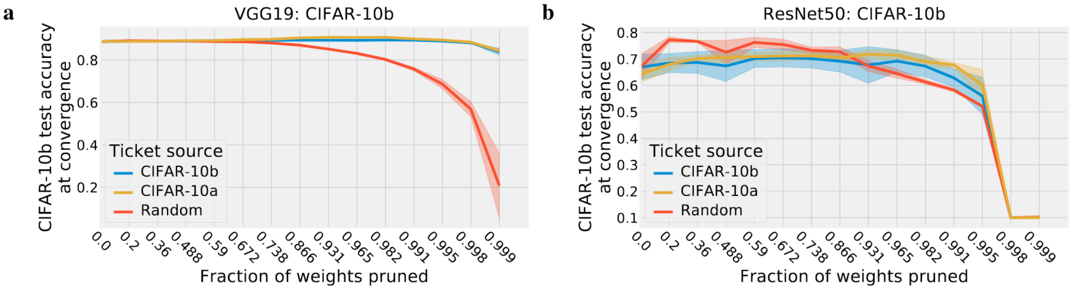


Figure 2: Morcos et al. [2019]

Morcos et al. [2019] conducted a series of empirical tests that would determine the cause of the utility of winning tickets. Three scenarios were chosen; transfer within the same data distribution, transfer across datasets and transfer across optimizers. Each experiment setup and their results are summarized below.

3.1 Transfer within the same data distribution

To test the plainest form of transfer, a dataset is divided into two(A and B). Winning tickets are generated on A and used on B.

In the experiments conducted by the authors of the paper mentioned above, CIFAR-10, Krizhevsky et al. [2009a] is divided into two halves, CIFAR-10a and CIFAR10b. The tickets generated using CIFAR10a were tested on the CIFAR10b dataset. Two architectures were used to drive this comparison, ResNet50 and VGG19. In addition, in all the experiments, a *random ticket* is also provided for comparison. A random ticket has randomly drawn initialization values as well as random masks.

Results The transferred tickets, in general, perform better compared to the random tickets. For VGG19, the performance of both the CIFAR10a and CIFAR10b tickets are good and the accuracy of random tickets taper off as the fraction of weights pruned increases.

However, this is not the case for ResNet50 where random tickets perform better than both CIFAR10a and CIFAR10b when the fraction of pruned weights is low. The performance for the transferred tickets only improves(compared to the random tickets) at a pruning fraction higher than 0.9. The authors speculate that this could be because of the winning tickets of the architecture being vulnerable to smaller datasets(CIFAR10a/b only contain 25,000 samples each).

3.2 Transfer across datasets

Transferring across datasets and architectures is the central theme of Morcos et al. [2019] which is summarised here. The six datasets used, all of them being natural image datasets are **Fashion-MNIST** Xiao et al. [2017], **SVHN** Netzer et al. [2011]. **CIFAR10**, **CIFAR100** Krizhevsky et al. [2009b], **ImageNet** Deng et al. [2009], **Places365** Zhou et al. [2017]. Figure 3 presents the results of the experiment conducted using the VGG19 architecture.

Random tickets are repeatedly outperformed by the transferred tickets as well as the dataset's original tickets. ImageNet consistently maintains a high accuracy(compared to other winning tickets) despite the pruning fraction, except when it comes to the Fashion-MNIST dataset(Figure 3), where the accuracy drops at 0.965. Also, the authors propose that transferred tickets provide *regularization against overfitting*. This statement is put forward because, VGG19, being a large network, overfits on smaller datasets(like Fashion-MNIST) in the experiment, resulting in low accuracy at low pruning rates. This does not happen in the case of transferred winning tickets. Also, looking at the results, it can be suggested that larger datasets/higher number of classes in the dataset provide better(more generalizable) winning tickets compared to smaller datasets (refer to winning tickets generated by ImageNet, Places365).

Further observations can be drawn from Figure 3 but for brevity, it can be concluded that **winning tickets that generalize well across datasets can be found**.

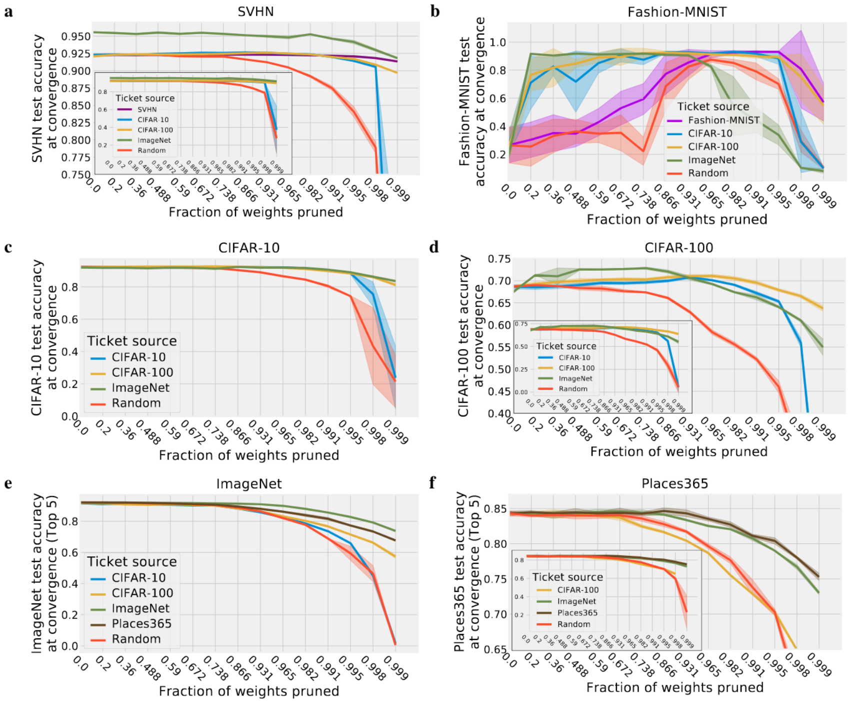


Figure 3: Morcos et al. [2019] (a) SVHN, (b) Fashion-MNIST, (c) CIFAR10, (d) CIFAR100, (e) ImageNet, (f) Places365

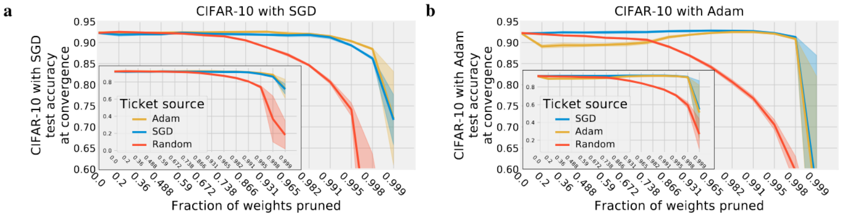


Figure 4: Morcos et al. [2019] (a) SGD, (b) Adam

3.3 Transfer across optimizers

Two optimizers(SGD and Adam) are used to conduct these experiments by Morcos et al. [2019]. Figure 4 shows the results when CIFAR10 is used as the dataset and VGG19 is the architecture. Winning tickets from both optimizers perform well when transferred. Interestingly, in figure 4b, random tickets perform better than the source and the transferred tickets for lower pruning fractions. Through a series of experiments, Morcos et al. [2019] demonstrate that winning tickets do learn inductive biases during training that can be generalized across datasets and optimizers.

3.4 Universality of winning tickets

The results and analysis provided by Morcos et al. [2019] suggest that it may be possible to find winning tickets that perform well across datasets, optimizers and architectures. Since the datasets used were from the images domain, a question arises, does this conclusion pan across multi-modalities ?

Further, recent advances in the field of Natural Language Processing(NLP) often involve increasing the number of parameters to produce better results and higher accuracy. Considering Large Language Models(LLM's) like BERT Devlin et al. [2018], which contain hundreds of millions of parameters, could transferable winning tickets be possible ?

Morcos et al. [2019] look at universality of winning tickets in terms of transferring them across architectures, optimizers and datasets in the natural image domain. Another way of looking at universality could be by testing if an NLP task performed across different datasets, using a sparse LLM could still provide a good accuracy. This is explored in section 4.1. But first, it is important to determine the affect that compressing a LLM has on its ability to transfer to new tasks. This is explored in the next section.

4 Pruning Large Language Models

Pre-trained models have revolutionized the field of NLP by making it easier to train a model for downstream tasks. They are generally trained on a language modelling task(like Masked Language Modelling) and then fine-tuned on labelled samples of a specific task(Name Entity Recognition(NER),Natural Language Inference(NLI) etc). These models, generally have higher accuracy compared to models trained specifically for a task. The high accuracy due to the overparameterization of LLM's comes at a cost since these dense structures require huge amounts of GPU memory. This makes it tricky to deploy them on light-weight applications.

Gordon et al. [2020] focus on **Compressing BERT**, an approach that would prune the number of parameters an LLM(in this case BERT) contains such that accuracy is not compromised, while the GPU usage remains as optimal as possible.

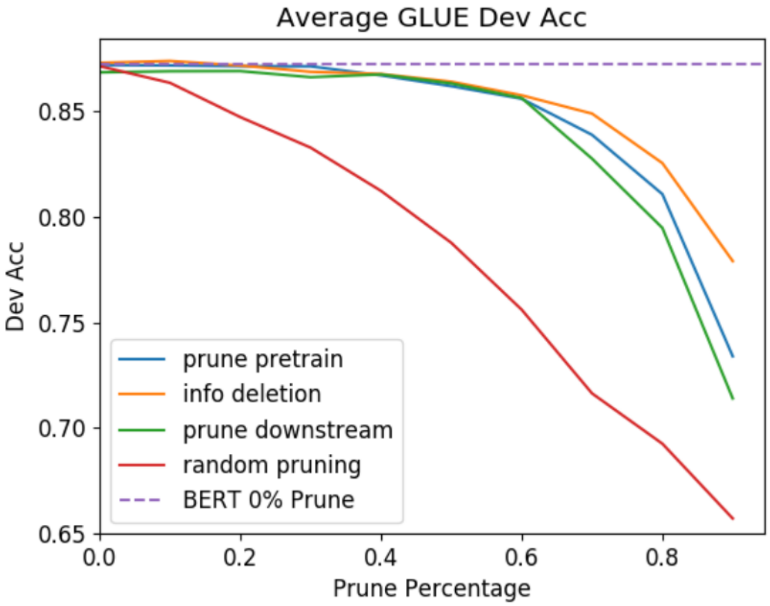


Figure 5: Gordon et al. [2020], These values are averaged over the 5 tasks. Orange - information deletion during pre-training, green - pruning after training on downstream task, red - random pruning(models are randomly pruned, pruning is not governed by magnitude of parameter), blue - best GLUE accuracy

Implementation and Setup: The authors of Gordon et al. [2020] experiment with BERT and use magnitude pruning to compress the model and then evaluate the model on the GLUE benchmark, Wang et al. [2018]. It is interesting that magnitude pruning was chosen and not the iterative version. Morcos et al. [2019] espouse the capability of iterative mangnitude pruning(IMP) over pruning a percentage of connections. It would be interesting to see the results of this experiment when IMP is the pruning technique.

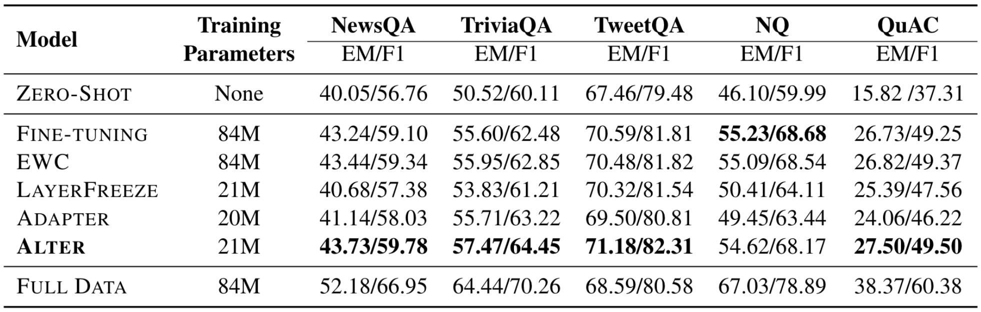
Additionally, the authors choose ”matrix-local” pruning over global pruning in contrast to Morcos et al. [2019], stating that matrix global pruning is used due to the fact that is more popular in the community.

The specifications of the local pruning method involve pruning each of the (stacked)key projections, query projections and value projections from all attention heads separately. Word embeddings are also pruned; further specifications are provided at Gordon et al. [2020].

The authors run the experiment on 5 out of the 9 tasks, which are MNLI(Multi-Genre Natural Language Inference), QQP(Quora Question Pairs), QNLI(Question NLI), SST-2(Semantic Textual Similarity Benchmark), CoLA(The Corpus of Linguistic Acceptability).

Results: The most significant result from this paper is that *30-40% of parameters can be dispensed with.* It does not matter if these weights are pruned before or after fine tuning. This conclusion is corroborated by Figure 5. When pruning beyond 40%, accuracy starts to taper off. This degradation in performance is not uniform; for each task, the performance devalues at different rates. Further, it is empirically shown that over-pruning BERT "deletes" useful information for different tasks.

It is now known that a certain fraction of BERT's weights can be discarded safely regardless of when the pruning is conducted. It would be interesting to look at how lottery tickets generated on BERT performs across different domains for a particular task, such as Reading Comprehension. Zhu et al. [2021] conduct experiments in this direction in the form of a domain adaptation task. The *source domain model* is BERT finetuned on the SQuAD Rajpurkar et al. [2016] dataset while the target domains are mentioned below in section 4.1.



Model	Training Parameters	NewsQA	TriviaQA	TweetQA	NQ	QuAC
		EM/F1	EM/F1	EM/F1	EM/F1	EM/F1
Zero-Shot	None	40.05/56.76	50.52/60.11	67.46/79.48	46.10/59.99	15.82 /37.31
Fine-tuning	84M	43.24/59.10	55.60/62.48	70.59/81.81	**55.23/68.68**	26.73/49.25
EWC	84M	43.44/59.34	55.95/62.85	70.48/81.82	55.09/68.54	26.82/49.37
LayerFreeze	21M	40.68/57.38	53.83/61.21	70.32/81.54	50.41/64.11	25.39/47.56
Adapter	20M	41.14/58.03	55.71/63.22	69.50/80.81	49.45/63.44	24.06/46.22
Alter	21M	**43.73/59.78**	**57.47/64.45**	**71.18/82.31**	54.62/68.17	**27.50/49.50**
Full Data	84M	52.18/66.95	64.44/70.26	68.59/80.58	67.03/78.89	38.37/60.38

Figure 6: Zhu et al. [2021], "EM" signifies exact match

4.1 Pruning BERT and the Reading Comprehension task

Zhu et al. [2021] discuss the effect pruning BERT(fine tuned on SQuAD Rajpurkar et al. [2016]) has on the Reading Comprehension task across different datasets.

In a Reading Comprehension task, generally, each question is associated with a passage and the answer to the question is a portion of the passage text. A popular example would be SQuAD, Rajpurkar et al. [2016]. Other reading comprehension datasets used in Zhu et al. [2021] are NewQA Trischler et al. [2016], TriviaQA Joshi et al. [2017], TweetQA Xiong et al. [2019], NaturalQuestions Kwiatkowski et al. [2019] and QuAC Choi et al. [2018].

Implementation and Setup: Zhu et al. [2021] introduce `ALTER` (Adaptable Lottery) and measure its performance against different baselines.

In the **Zero-Shot** baseline no input regarding the task is provided to the source domain model, **Fine-tuning** fine tunes the entire source domain model on the target domain. **Layer Freeze** only fine tunes the top few layers and freezes the rest. **Elastic Weight Consolidation(EWC)** Kirkpatrick et al. [2017] is an algorithm that forces parameters to stay close to their true value and lastly, **Adapter** Houlsby et al. [2019] is a kind of model that allows parameter sharing between different tasks through the use of adapter modules which only adds a few trainable parameters per task. `FULL DATA` is the model that uses the full training set without adaptation.

The implementation can be summarized as follows: The source domain model is first pruned using a *gradual* magnitude pruning algorithm instead of IMP. Self-attention attribution Hao et al. [2021] is used to prune parameters while keeping in mind that parameters from the important attention heads are preserved.

After pruning, a lottery sub-network is obtained, which is then used to fine tune on the target domain. The parameters that are not in the lottery sub-network are frozen, so they are still part of the training but not updated.

The results for this experiment are specified in Figure 6. **ALTER performs better on 4 out of the 5 target domains**. The authors also conclude that the attention to the fact that parameters of important attention heads should receive special consideration does indeed improve the quality of the lottery sub-networks and improves the performance.

5 Conclusion

The Lottery Ticket Hypothesis is a breakthrough in the sense that future work on LLM's must focus on optimising the size of models. The exponential growth in the size of models improves accuracy but it is imperative that GPU usage as well as the efficient deployment of these large scale models is put on the forefront. The suspected universality of LLM's across tasks provide hope that a lucky lottery sub-network is enough to train domain/task-specific models. This report sheds light on the topic of pruning in LLM's, specially focusing on the different ways(within a task, between tasks and so on) that winning tickets are ”universal” in the context of LLM's. Current research seems to focus on BERT as the preferred LLM for these experimentations, it would be interesting to see if the conclusions by the authors cited in this paper are valid across LLM architectures.

References

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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey 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. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. 2018. doi: 10.48550/ARXIV.1803.03635. URL https://arxiv.org/abs/1803.03635.

Mitchell A Gordon, Kevin Duh, and Nicholas Andrews. Compressing bert: Studying the effects of weight pruning on transfer learning. *arXiv preprint arXiv:2002.08307*, 2020.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009a.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009b.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Robert Tjarko Lange. The lottery ticket hypothesis: A survey. *https://roberttlange.github.io/year-archive/posts/2020/06/lottery-ticket-hypothesis/*, 2020. URL https://roberttlange.github.io/posts/2020/06/lottery-ticket-hypothesis/.

Ari S. Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers, 2019. URL https://arxiv.org/abs/1906.02773.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016.

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

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. Tweetqa: A social media focused question answering dataset. *arXiv preprint arXiv:1907.06292*, 2019.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464, 2017.

Haichao Zhu, Zekun Wang, Heng Zhang, Ming Liu, Sendong Zhao, and Bing Qin. Less is more: Domain adaptation with lottery ticket for reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1102–1113, 2021.