On the weak link between importance and prunability of attention heads

Aakriti Budhraja Madhura Pande Preksha Nema
Pratyush Kumar Mitesh M. Khapra
Robert Bosch Centre for Data Science and Artificial Intelligence (RBC-DSAI)
IIT Madras, India
{abudhra,mpande,preksha,pratyush,miteshk}@cse.iitm.ac.in

Abstract

Given the success of Transformer-based models, two directions of study have emerged: interpreting role of individual attention heads and down-sizing the models for efficiency. Our work straddles these two streams: We analyse the importance of basing pruning strategies on the interpreted role of the attention heads. We evaluate this on Transformer and BERT models on multiple NLP tasks. Firstly, we find that a large fraction of the attention heads can be randomly pruned with limited effect on accuracy. Secondly, for Transformers, we find no advantage in pruning attention heads identified to be important based on existing studies that relate importance to the location of a head. On the BERT model too we find no preference for top or bottom layers, though the latter are reported to have higher importance. However, strategies that avoid pruning middle layers and consecutive layers perform better. Finally, during fine-tuning the compensation for pruned attention heads is roughly equally distributed across the un-pruned heads. Our results thus suggest that interpretation of attention heads does not strongly inform pruning.

1 Introduction

The acclaimed success of Transformer-based models across NLP tasks has been followed by two important directions of research. In the first direction, interpretability studies aim to understand how these models work. Given that multi-headed attention is an important feature of these models, researchers have focused on attention heads as the units of interpretation. These studies comment on the role of each attention head and the relation between a head's position and its significance (Clark et al., 2019; Michel et al., 2019; Voita et al., 2019b,a; Liu et al., 2019; Belinkov et al., 2017). These studies show that certain heads are more important based on (i) their position in the network (top, middle, bottom), or (ii) the component to which they belong (encoder self-attention, decoder self-attention, encoder-decoder cross attention), or (iii) the functional role they play (e.g., syntactic/semantic).

In the other major direction, these large Transformer-based models have been down-sized to be more time and space efficient. Different methods for down-sizing have been studied such as pruning (McCarley, 2019; Gordon et al., 2020; Sajjad et al., 2020), distillation (Sanh et al., 2019; Liu et al., 2019; Jiao et al., 2019), weight quantization (Zafrir et al., 2019; Shen et al., 2019), and weight factorization and parameter sharing (Lan et al., 2019). Pruning techniques have been particularly successful in reinforcing the folk-lore that these models are highly over-parameterized. These pruning methods prune parameters based on magnitude (Gordon et al., 2020), importance (McCarley, 2019) or layer-wise (Sajjad et al., 2020).

In this paper, we straddle these two directions of work by asking the following question: *Can we randomly prune heads, thus completely ignoring any notion of importance of heads?* To answer this, we systematically study the effect of randomly pruning specific subsets of attention heads on the accuracy on different tasks. Across experiments, we modify the random sampling to vary the percentage of heads pruned and their location in the network (components and layers).

We evaluate these experiments both on the Transformer and BERT models. Our results show that a large fraction of attention heads can be pruned randomly: 75% of the attention heads of Transformer can be randomly pruned with a drop of less than 1 BLEU point on NMT tasks. Similarly, half of the attention heads of BERT can be randomly pruned with an average drop in accuracy of less

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3230–3235,
November 16–20, 2020.

than 1% across a chosen set of GLUE tasks[1]. Significantly for Transformers, we find no evidence for pruning methods preferring specific attention heads based on their location; even when the locations are chosen to match attention heads identified to be more important in existing studies. Similarly on the BERT model, pruning top and bottom layers do not show significant difference, even though existing studies attribute higher importance to the latter (Sajjad et al., 2020). However, we identify a preference to avoid pruning the middle layers and consecutive layers. Lastly, we check if during fine-tuning certain heads compensate more for the pruned heads. If so, such heads would perhaps be more important. However, we find no such evidence. In particular, during fine-tuning, the un-pruned heads change similarly across most pruning configurations. Overall, our experiments suggest that interpretation of attention heads does not strongly inform pruning.

The rest of the paper is organized as follows: Section 2 mentions about the models and the datasets used for this work followed by Section 3 which provides details of the experimental process. This section reports results on both Transformer and BERT models. We summarize our work in Section 4.

2 Models and Datasets

2.1 Multi-headed Self Attention

In each multi-headed attention layer we have multiple attention heads which transform the representation of inputs of a given sequence of tokens. Given the d_v dimensional representation of T tokens as $X \in \Re^{T \times d_v}$, the output of multi-headed self attention with N attention heads is given by

$$\text{Concat}_{i=1}^{N} \left(\text{softmax} \left(\frac{(XW_i^q)(XW_i^k)^T}{\sqrt{d_k}} \right) XW_i^v \right), \quad (1)$$

where $W_i^k, W_i^q, W_i^v \in \Re^{d_v \times d_k}$ are parameters of the i-th attention head.

2.2 Transformers

We use the Transformer-Base model (Vaswani et al., 2017) which has 6 layers each in the three components: encoder self-attention (ES), encoder-decoder cross-attention (ED), and decoder self-attention (DS). In each layer of each of the three components, we have 8 attention heads, totalling to $3 \times 6 \times 8 = 144$ attention heads. We train the models with 2.5 million sentence pairs each from the WMT'14 English-Russian (EN-RU) and English-German (EN-DE) datasets. We report BLEU scores on WMT's *newstest2014*. We use Adam optimizer (Kingma and Ba, 2014) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.997$, and $\epsilon = 10^{-9}$. We vary the learning rate according to the formula described in Vaswani et al. (2017) with *warmup_steps* = 16k. We use large batch sizes of 32k and 25k for EN-RU and EN-DE, respectively, as it has been established that large batch sizes are inherent to the performance of Transformers (Popel and Bojar, 2018; Voita et al., 2019b). We achieve effectively large batch sizes using the technique of gradient accumulation on single NVIDIA V100 and 1080Ti GPUs.

2.3 BERT

In all experiments involving BERT, we use the BERT Base-uncased model (Devlin et al., 2018). It has 12 layers and each layer contains 12 attention heads, summing to 144 attention heads. We fine-tune and evaluate the pre-trained model[2] on sentence entailment task MNLI-M, the question similarity task QQP, the question-answering task QNLI, and the movie review task SST-2 from the GLUE Benchmark (Wang et al., 2018). We report accuracies on the official development sets of the considered GLUE tasks. For each of the four GLUE tasks, namely MNLI-M, QQP, QNLI and SST-2, we tried combinations of batch size and learning rate from $\{8, 16, 32, 64, 128\}$ and $\{2, 3, 4, 5\} \times 10^{-5}$ respectively and selected the best performing configuration. The exact hyperparameters used for each of the tasks have been made available with the code released[3]. Each BERT experiment was run on a single Cloud TPU (v2-8).

3 Experiments

3.1 Experimental Process

In all the experiments, we perform random pruning where a subset of attention heads chosen by random sampling are zeroed out. Formally, each attention head is assigned a weight ξ which is 0 if the head is pruned and 1 otherwise. Then, the output of an attention layer is given by

$$\text{Concat}_{i=1}^{N} \left(\xi_i \text{softmax} \left(\frac{(XW_i^q)(XW_i^k)^T}{\sqrt{d_k}} \right) XW_i^v \right) \quad (2)$$

After pruning, we fine-tune the Transformer model for 30 epochs and the BERT model for 10 epochs.

[1] We avoid WNLI, RTE, MRPC, STS-B, CoLA as the results on these datasets tend to be noisy and unstable as reported in (Gordon et al., 2020; Sajjad et al., 2020)

[2] https://github.com/google-research/bert

[3] https://github.com/iitmnlp/head_importance_and_pruning

Since the values ξ are randomly sampled, in each experiment we report the average of three different samplings of ξ. The standard deviations are 0.668% and 0.778% of the reported average values for Transformer and BERT respectively.

3.2 Experimental Results on Transformers

Varying Pruning Percentage. We randomly prune attention heads across all components and layers varying the percentage of pruning from 25% to 87% (Table 1). We observed that in the case of extreme pruning, i.e., keeping just one head in each layer of each of the three components (which corresponds to a pruning percentage of 87%), the drop in BLEU was 1.62 (EN-RU) and 1.03 (EN-DE) as can be seen from Table 1. Across both EN-RU and EN-DE tasks, 60% of the attention heads can be pruned with a maximum drop in BLEU score by only 0.15. As can be observed from Figure 1, the drop is sharper as we increase the pruning percentage beyond 60%.

% Pruning	EN-RU	EN-DE
0 (Baseline)	29.09	27.95
25	29.59 (+0.50)	28.19 (+0.24)
35	29.29 (+0.20)	27.94 (-0.01)
50	29.38 (+0.29)	28.02 (+0.07)
55	29.00 (-0.09)	28.24 (+0.29)
60	28.94 (-0.15)	27.88 (-0.07)
75	28.22 (-0.87)	27.49 (-0.46)
81	27.97 (-1.12)	26.80 (-1.15)
87	27.47 (-1.62)	26.92 (-1.03)

Table 1: BLEU scores for Transformer on EN-RU and EN-DE datasets when subject to varying pruning percentages. Difference from the baseline score is indicated in brackets.

Pruning based on Layer Numbers. Voita et al. (2019b) identify that attention heads in specific layers of the Transformer – lower layers of Self-Attention components, i.e., Encoder-Self (ES) and Decoder-Self (DS), and higher layers of Encoder-Decoder cross attention (ED) – are more important. We evaluate the correspondence of this importance to pruning. We choose 5 pruning percentages from 25% to 75% and in each case two pruning configurations: One where the heads considered important are retained and the other where the important heads are pruned. The configurations and the corresponding BLEU scores on the EN-RU dataset are shown in Table 2 where each configuration is specified as a string. For example, the string 777322 indicates that 7 heads each were retained in the first three layers, 3 in the fourth layer and 2 each in the last two layers. For each pruning percentage, the first row corresponds to the configuration in which heads considered important (Voita et al., 2019b) were retained and the second row corresponds to the adversarial configuration in which heads considered important were pruned. We identify no preference in pruning as for each pruning percentage the performance of both configurations is very similar.

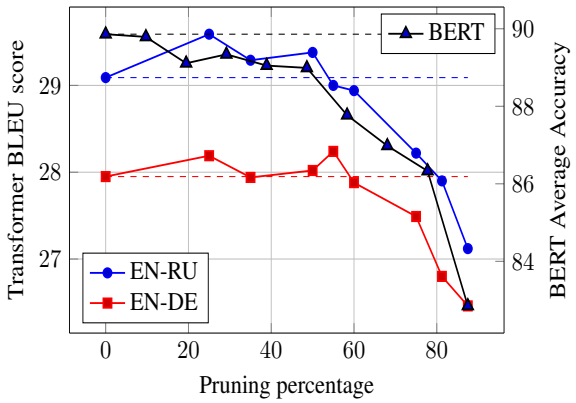

Figure 1: Effect of random pruning on the performance of Transformer and BERT for various pruning percentages.

%	Configuration ES	Configuration ED	Configuration DS	BLEU Scores
0	888888	888888	888888	29.09
25	888444	444888	888444	29.62 (+0.53)
	444888	888444	444888	29.43 (+0.34)
40	777322	233777	777332	29.17 (+0.08)
	223777	777332	233777	29.57 (+0.48)
50	666222	222666	666222	29.01 (-0.08)
	222666	666222	222666	29.35 (+0.26)
60	555211	112555	555211	28.99 (-0.10)
	112555	555211	112555	28.78 (-0.31)
75	333111	111333	333111	28.48 (-0.61)
	111333	333111	111333	28.35 (-0.74)

Table 2: BLEU scores for different pruning configurations of Transformer. Every row has 2 configurations: first, where the important heads are retained, and second, where the important heads are pruned.

Pruning Based on Component. Some studies show that heads in the ED component are most important while those in the ES module are least important (Voita et al., 2019b). We choose 4 different pruning percentages and in each case consider three configurations where the number of attention heads is least in one chosen component (ES, ED, DS). The configurations and corresponding BLEU scores on the EN-RU dataset are shown in Table 3.

Pruning %	Configuration	BLEU Score
Baseline	(48,48,48)	29.09
48%	(14,31,30)	28.96 (-0.13)
	(31,14,30)	29.00 (-0.09)
	(30,31,14)	29.13 (+0.04)
60%	(12,21,25)	28.48 (-0.61)
	(21,12,25)	28.78 (-0.31)
	(25,21,12)	28.48 (-0.61)
75%	(8,13,15)	27.95 (-1.14)
	(13,8,15)	27.96 (-1.13)
	(15,13,8)	28.04 (-1.05)
82%	(5,9,12)	27.24 (-1.85)
	(9,5,12)	26.95 (-2.14)
	(12,9,5)	27.83 (-1.26)

Table 3: BLEU scores for different pruning configurations of Transformer specified by the triple denoting the number of heads retained in the Encoder-Self, Encoder-Decoder, and Decoder-Self attention components.

We identify no consistent preference in the pruning strategy: In the 4 cases considered, each of the 3 configurations has the highest BLEU score in at least one case. Note that we chose the number of heads in each layer (14, 31, etc) to be consistent with those used in (Voita et al., 2019b).

3.3 Experimental Results on BERT

% Pruning	MNLI-M	QQP	QNLI	SST-2
0	83.69	91.22	91.66	92.88
10	83.70	91.39	91.60	92.48
20	82.80	91.09	90.33	92.25
30	82.87	91.19	90.84	92.48
40	82.48	91.05	90.40	92.27
50	83.02	90.90	90.04	92.00
60	81.35	90.56	87.31	91.89
70	80.40	89.83	86.85	90.86
80	78.93	90.03	86.40	89.96
90	75.08	87.44	81.80	87.11

Table 4: Performance of random pruning on BERT for different pruning percentages. The accuracies are reported on the official GLUE development datasets.

Varying Pruning Percentage. We vary the pruning percentage from 10 to 90% and report the accuracy on the 4 GLUE tasks: MNLI-M, QQP, QNLI, and SST-2 (Table 4). We observe that half of the attention heads can be pruned with an average accuracy drop of under 1%. As shown in Figure 1, beyond 50% pruning, the accuracy drop is sharper.

Pruning based on Layer Numbers. To identify any preference to pruning heads in specific layers, we consider several configurations as shown in Table 5, where we prune a subset of layers entirely, i.e. we prune all the attention heads of particular layers. When all the self-attention heads of a layer *l* are pruned, only the feed-forward network of that layer will be active whose input will just be the output from the previous layer *l-1*.

Layers Pruned	MNLI-M	QQP	QNLI	SST-2
0 (Baseline)	83.69	91.22	91.66	92.88
Top 1	82.95	91.33	91.48	91.85
Bottom 1	83.65	91.42	91.17	93.11
Top 3	82.58	90.85	89.2	92.31
Bottom 3	83.36	90.95	90.88	92.54
Top 6	80.98	90.52	87.44	90.02
Bottom 6	79.29	90.17	87.40	91.05
Top 8	77.59	89.34	85.08	88.53
Bottom 8	78.07	89.67	84.22	87.95
Top 1, Bottom 1	83.33	91.23	90.70	92.88
Middle 2	83.60	91.08	90.80	91.74
Top 2, Bottom 2	82.41	91.11	90.48	92.20
Middle 4	81.84	90.87	86.14	90.94
Top 3, Bottom 3	81.72	90.67	88.30	92.31
Middle 6	80.08	90.49	87.07	87.84
Top 4, Bottom 4	79.47	89.57	86.01	90.36
Middle 8	78.88	89.55	83.67	88.87

Table 5: Accuracy on GLUE tasks for multiple layer-wise pruned configurations of BERT.

Bottom layers of BERT have been identified to model word morphology (Liu et al., 2019; Belinkov et al., 2017) and are considered to be important (Sajjad et al., 2020). Further, recent work has identified high cosine-similarity between output vectors of the top layers, indicating reduced importance of top layers (Goyal et al., 2020). We relate these studies to pruning by comparing the pruning of the same number of top and bottom layers (rows 2-9 in Table 5). Amongst the four cases, two cases each favor pruning top layers and bottom layers, revealing no preference in pruning.

The middle layers in BERT have been shown to have specific characteristics of higher attention entropy and greater attention to specific tokens (Clark et al., 2019). We thus considered configurations where we compare pruning top and bottom layers against pruning middle layers (last eight rows of Table 5). The results indicate a clear preference: In 14 out of 16 cases, pruning the middle layers performs worse that pruning equal number of layers distributed among top/bottom layers. Indeed, we incur an additional over 2% average drop in accuracy for QNLI and SST-2 tasks, indicating a task-specific sensitivity to pruning middle layers.

Recent work has identified that consecutive lay-

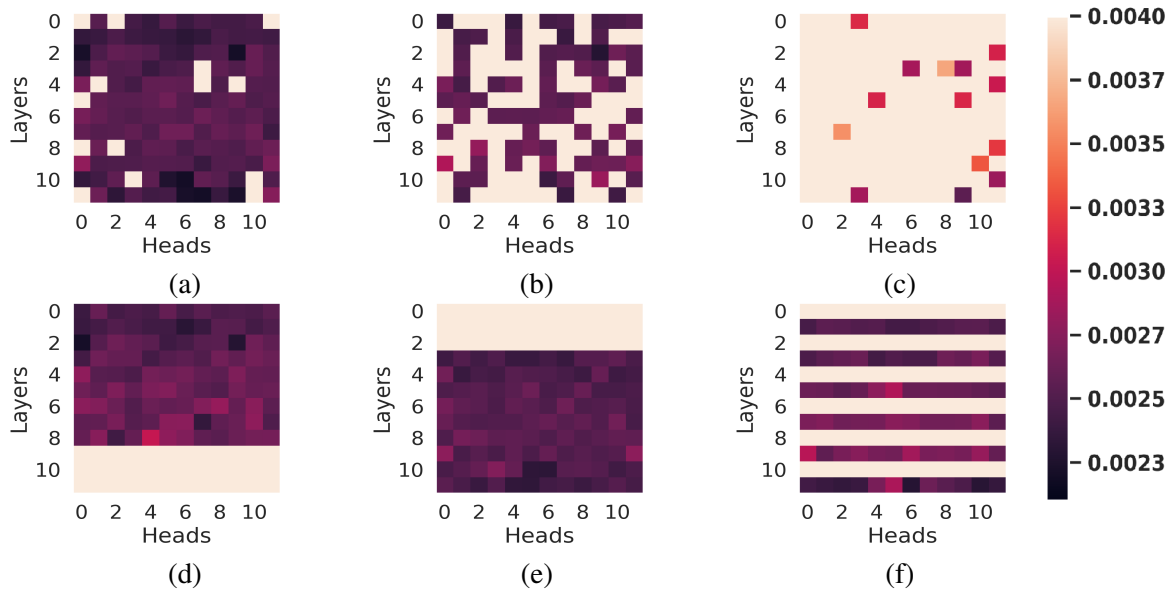

Figure 2: Head-wise average magnitude change of weights during fine-tuning for the following pruning configurations of BERT for the MNLI-M task: (a) 10% pruned (b) 50% pruned (c) 90% pruned (d) Top three layers pruned (e) Bottom three layers pruned (f) Alternate layers pruned.

ers of BERT have similar functionality (Lan et al., 2019). To study this, we considered configurations where six even and odd alternate layers are pruned and compare it with other strategies of pruning 50% layers of BERT (Table 6). We observe that the odd configuration performs better than the Top 6 and Bottom 6 configurations, indicating a preference to avoid pruning of consecutive layers.

Layers Pruned	MNLI-M	QQP	QNLI	SST-2
Top 6	80.98	90.52	87.44	90.02
Bottom 6	79.29	90.17	87.40	91.05
Even 6	81.54	90.74	86.39	90.36
Odd 6	81.95	90.58	90.18	92.20
Top 3, Bottom 3	81.72	90.67	88.30	92.31
Middle 6	80.08	90.49	87.07	87.84

Table 6: Accuracy on GLUE tasks when half of the layers of BERT are pruned. Pruning odd numbered layers retains the maximal accuracy across most of the tasks.

Effect of Fine-Tuning. Recent studies (Kovaleva et al., 2019; Houlsby et al., 2019) have reported that when fine-tuning BERT for specific tasks, the top layers change much more than the lower layers. We now evaluate this for fine-tuning after pruning. In Figure 2, we plot the average change in magnitude of parameters for different attention heads (W^q, W^k, W^v in Equation 1) for the MNLI-M task. We observe no spatial patterns in the parameter changes or with respect to relative distance from pruned heads. In particular, for all experiments in Table 5 and 6, the average change in attention parameters for any two layers differs by less than 10%. This shows that the compensation for pruned attention heads is roughly equally distributed across the unpruned heads.

4 Conclusion

We systematically studied the effect of pruning attention heads in Transformer and BERT models. We confirmed the general expectation that a large number of attention heads can be pruned with limited impact on performance. For Transformers we observed no preference for pruning attention heads which have been identified as important in interpretability studies. Similarly, for BERT we found no preference between pruning top and bottom layers. However, pruning middle layers and consecutive layers led to a larger drop in accuracy. We also observe that the recovery during fine-tuning was uniformly distributed across attention heads. We conclude that there is often no direct entailment between importance of an attention head as characterised in several recent studies, and low prunability of the respective head using random pruning.

Acknowledgements

We thank Amazon Web Services for their support with the NVIDIA GPUs. We also thank Google for the free TPU credits under their TFRC program,

and for supporting Preksha Nema through their Google Ph.D. India Fellowship program. We also thank the Department of Computer Science and Engineering as well as the Robert Bosch Centre for Data Science and Artificial Intelligence (RBC-DSAI), IIT Madras for providing us with all the resources that made this work possible.

References

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

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*.

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

Saurabh Goyal, Anamitra Roy Choudhary, Venkatesan Chakaravarthy, Saurabh ManishRaje, Yogish Sabharwal, and Ashish Verma. 2020. Power-bert: Accelerating bert inference for classification tasks. *arXiv preprint arXiv:2001.08950*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*.

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*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

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*.

Linqing Liu, Huan Wang, Jimmy Lin, Richard Socher, and Caiming Xiong. 2019. Attentive student meets multi-task teacher: Improved knowledge distillation for pretrained models. *arXiv preprint arXiv:1911.03588*.

J Scott McCarley. 2019. Pruning a bert-based question answering model. *arXiv preprint arXiv:1910.06360*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2020. Poor man's bert: Smaller and faster transformer models. *arXiv preprint arXiv:2004.03844*.

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*.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2019. Q-bert: Hessian based ultra low precision quantization of bert. *arXiv preprint arXiv:1909.05840*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. *arXiv preprint arXiv:1909.01380*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

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*.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. *arXiv preprint arXiv:1910.06188*.