

Poor Man’s BERT: Smaller and Faster Transformer Models

Hassan Sajjad

Fahim Dalvi

Nadir Durrani

Preslav Nakov

Qatar Computing Research Institute, HBKU Research Complex, Doha 5825, Qatar
 {hsajjad, faimaduddin, ndurrani, pnakov}@hbku.edu.qa

Abstract

The ongoing neural revolution in Natural Language Processing has recently been dominated by large-scale pre-trained Transformer models, where size does matter: it has been shown that the number of parameters in such a model is typically positively correlated with its performance. Naturally, this situation has unleashed a race for ever larger models, many of which, including the large versions of popular models such as BERT, XLNet, and RoBERTa, are now out of reach for researchers and practitioners without large-memory GPUs/TPUs. To address this issue, we explore a number of memory-light model reduction strategies that do not require model pre-training from scratch. The experimental results show that **we are able to prune BERT, RoBERTa and XLNet models by up to 40%, while maintaining up to 98% of their original performance**. We also show that **our pruned models are on par with DistilBERT in terms of both model size and performance**. Finally, **our pruning strategies enable interesting comparative analysis between BERT and XLNet**.¹

1 Introduction

Pre-trained neural language models have achieved state-of-the-art performance on natural language processing tasks and have been adopted as feature extractors for solving downstream tasks such as question answering, natural language inference, and sentiment analysis. The current state-of-the-art Transformer-based pre-trained models consist of dozens of layers and millions of parameters. While deeper and wider models yield better performance, they also need large GPU/TPU memory. For example, BERT-large (Devlin et al., 2019) is 24-layered

and has 335 million parameters, which requires at least 24 GB of GPU memory. The large memory requirement of these models limit their applicability, e.g., they cannot be run on small hand-held devices. Additionally, these models are slow at inference time, which makes them impractical to deploy in real-time scenarios.

Recently, several methods have been proposed to reduce the size of pre-trained models, particularly, BERT. Some notable approaches include strategies to prune parts of the network after training (Michel et al., 2019a; Voita et al., 2019b; McCarley, 2019), reduction through weight factorization and sharing (Lan et al., 2019), compression through knowledge-distillation (Sanh et al., 2019) and quantization (Zafir et al., 2019; Shen et al., 2019). Our work falls under the class of pruning methods.

The central argument governing pruning methods is that deep neural models are over-parameterized and that not all parameters are strictly needed, especially at inference time. For example, recent research has shown that some attention heads can be removed at test time without significantly impacting performance (Michel et al., 2019b; Voita et al., 2019b). Our work builds on similar observations, but instead we question whether it is necessary to use all layers of a pre-trained model in downstream tasks. To answer this, we propose straight-forward strategies to drop some layers from the neural network.

Our proposed strategies are motivated by recent findings in representation learning. For example, Voita et al. (2019a) showed that the top layers of the pretrained models are inclined towards the pre-trained objective function. Michel et al. (2019b) commented on the over-parameterization and the redundancy in the pre-trained model, leading us to hypothesize that adjacent layers might preserve redundant information. Liu et al. (2019b) reported that certain linguistic tasks, such as word morphol-

¹The code is available at <https://github.com/hsajjad/transformers>. We only change the run_glue.py file. Rest of the code is same as that of the Transformer library.

ogy, are learned at the lower layers of the network, whereas higher-level phenomena are learned at middle and higher layers in the network.

In the light of these findings, we explore several strategies to drop layers from a pre-trained Transformer model. More specifically, we drop top layers, bottom layers, middle layers, alternate layers, or layers that contribute least in the network. We first remove a set of layers from the pre-trained model based on a strategy, and then we fine-tune the reduced model towards downstream tasks.

We apply our strategies across three state-of-the-art pre-trained models, BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019c), evaluating their performance against GLUE tasks (Wang et al., 2018). We also compare our results to DistilBERT (Sanh et al., 2019) and DistilRoBERTa, distilled versions of BERT and RoBERTa respectively, trained using the student-teacher paradigm. Our findings and contributions are summarized as follows:

- We present practical strategies for reducing the size of a pre-trained multi-layer model, while preserving up to 98.2% of its original performance.
- Our reduced models perform on par with distilled version of the pre-trained BERT and RoBERTa models in terms of accuracy, model size and inference speed, without requiring costly pre-training from scratch. This raises questions about the effectiveness of the current applications of knowledge distillation methods on pre-trained models and encourages further research in this direction.
- Our strategies are complementary to other distillation methods such as quantization and knowledge distillation. We further apply our best dropping strategy to DistilBERT, yielding an even smaller model with minimal performance degradation.
- We made interesting discoveries about the differences in XLNet and BERT. For example XLNet learns sequence-level knowledge earlier in the network and is thus more robust to pruning. To the best of our knowledge, this is the first work that provides a comparative analysis between these two models.
- Our setup enables practitioners to control the trade-off between model parameters and accuracy.

The remainder of this paper is organized as follows: Section 2 presents our layer-dropping strategies. Section 3 describes our experimental setup. Section 4 presents the results. Section 5 offers an in-depth discussion. Section 6 summarizes the related work. Finally, Section 7 concludes.

2 Layer-dropping Strategies

Consider a pre-trained language model M with an embedding layer E_0 and L encoder layers: $\{l_1, l_2, \dots, l_L\}$. We explore five strategies to drop encoder layers from model M .

2.1 Top-Layer Dropping

Previous research has shown that the top layers of the pre-trained models are specialized towards the underlying objective function (Voita et al., 2019a). Based on this observation, we hypothesize that the higher layers of the pre-trained model may not be important when fine-tuning towards the objective of a downstream task.

In this strategy, we drop top K layers from the model. The output of layer l_{L-K} serves as the last layer of the reduced network. Then, a task-specific layer is added on top of this layer to perform task-specific fine-tuning. Figure 1 shows an example with dropping top 4 and 6 layers.

2.2 Alternate Dropping

Deep neural networks are characterized by redundancy of information across the network (Tang et al., 2019). We hypothesize that neighbouring layers preserve similar information and it might be safe to drop alternate layers.

In this strategy, we drop N alternating odd or even layers from top to bottom of the network. For example for a 12-layer model with $K = 4$, we consider two sets of alternate layers: *Odd-alternate Dropping* – $\{5, 7, 9, 11\}$ and *Even-alternate Dropping* – $\{6, 8, 10, 12\}$ (See Figure 1 for illustration). When dropping an in-between layer l_i , the output of the previous layer l_{i-1} becomes the input of the next layer l_{i+1} , causing a mismatch in the expected input to l_{i+1} . However, we assume that during task-specific fine-tuning, the model will recover from this discrepancy.

2.3 Contribution-Based Dropping

Our next strategy is based on the idea that a layer contributing below a certain threshold might be a good candidate for dropping. We define contribution of a layer l_i in terms of the cosine similarity

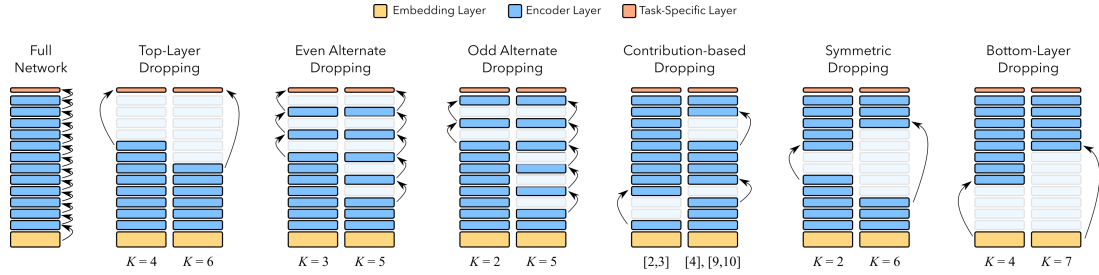


Figure 1: Layer-dropping Strategies

between its input and its output representations. A layer l_i with a high similarity (above a certain threshold) indicates that its output has not changed much from its input, and therefore it can be dropped from the network.

More concretely, in the forward pass, we calculate the cosine similarity between the representation of the sentence token (CLS) before and after each layer. We average the similarity scores of each layer over the development set, and select layers that have an average similarity above a certain threshold for dropping. This *contribution-based strategy* can be seen as a principled variation of *alternate dropping*.

2.4 Symmetric Dropping

It is possible that both lower and higher layers are more important than the middle layers in a model. Thus, we further experiment with *Symmetric dropping*, where we keep the top and the bottom X layers, and drop K middle layers, where $2X + K = L$. For example, in a 12-layer model, if $K = 6$, we retain three top and three bottom layers, dropping layers 4–9. The output of layer 3 would then serve as an input to layer 10 as can be seen in Figure 1.

2.5 Bottom-Layer Dropping

Previous work (Belinkov et al., 2017) has shown that lower layers model local interaction between words and word pieces (which is important for morphology and part of speech) thus providing essential input to higher-level layers. It is probably not a good idea to remove these lower layers. Here, we do it anyway for the sake of completeness of our experiments.

We remove the bottom K layers of the model. The output of the embedding layer l_0 serves as an input to layer l_{K+1} of the original model.

3 Experimental Setup

Datasets We evaluated our strategies on General Language Understanding Evaluation (GLUE) tasks (Wang et al., 2018). It involves a variety of language understanding tasks and now serves as a defacto standard to evaluate pre-trained language models. More specifically, we evaluated on the following tasks: SST-2 for sentiment analysis with the Stanford sentiment treebank (Socher et al., 2013), MNLI for natural language inference (Williams et al., 2018), QNLI for Question NLI (Rajpurkar et al., 2016), QQP for Quora Question Pairs,² RTE for recognizing textual entailment (Bentivogli et al., 2009), MRPC for Microsoft Research paraphrase corpus (Dolan and Brockett, 2005), and STS-B for the semantic textual similarity benchmark (Cer et al., 2017). We did not evaluate on WNLI, due to the irregularities in its dataset, as also found by others.³ We also excluded CoLA due to large variance and unstable results across many fine-tuning runs.

Models We experimented with BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019c) and DistilBERT (Sanh et al., 2019) using the `transformers` library (Wolf et al., 2019). We could not perform experiments with BERT-large or XLNet-large due to memory limitations⁴ and experiment only with the base models. However, our strategies are independent of any specific model and are straightforward to apply to models of any depth. We include DistilBERT for a comparative baseline, and to explore whether an already distilled model can be pruned further effectively.

²<http://data.quora.com/First-Quora-Dataset-Release-Question-Pairs>
³<http://gluebenchmark.com/faq>

⁴In order to fit large models in our TitanX 12GB GPU cards, we tried to reduce the batch size, but this yielded poor performance, as previously reported by the BERT team <https://github.com/google-research/bert#out-of-memory-issues>.

End-to-End Procedure Given a pre-trained model, we dropped layers using one of the strategies described in Section 2. We then performed task-specific fine-tuning using GLUE training sets for three epochs as prescribed by Devlin et al. (2019) and evaluated on the official devsets. We experimented with using more epochs, especially for dropping strategies that exclude in-between layers, in order to let the weight matrix adapt to the changes. However, we did not see any benefit in going beyond three epochs.

4 Evaluation Results

We experimented with dropping K number of layers where $K = 2, 4, 6$ in BERT and XLNet, and $K = 1, 2, 3$ in DistilBERT (a 6-layer model). As an example, for $K = 2$ on a 12-layer model, we drop the following layers: top strategy – $\{11, 12\}$; bottom strategy – $\{1, 2\}$; even-alternate – $\{10, 12\}$; odd-alternate – $\{9, 11\}$; symmetric – $\{6, 7\}$. In the contribution-based strategy, the dropping of layers is dependent on a similarity threshold. We calculate the similarity between input and output of each layer and remove layers for which the similarity is above the threshold values of 0.95, 0.925 and 0.9. These values were chosen empirically. A threshold value below 0.9 or above 0.95 resulted in either more than half of the network being considered as similar, or none of the layers to be similar. We discuss this approach further in Section 5.

4.1 Comparing Strategies

Figure 2 presents average classification performance of BERT and XLNet using our proposed strategies. Our results show that *Top-layer dropping* consistently outperforms other strategies when dropping 4 and 6 layers. We dropped half of the top layers (yellow bar in the top strategy) with an average loss of only 2.91 and 1.81 points for BERT and XLNet respectively. Similarly, dropping one-third of the network (i.e. 4 layers) resulted in a drop of only 2 points and 0.23 points for BERT and XLNet respectively (blue bar in the top strategy). The Bottom-layer dropping strategy performed the worst across all models, reflecting that it is more damaging to remove information from the lower layers of the network.

The *Odd-alternate dropping* strategy gave better results than the top at $K = 2$ (blue bars in the Odd-alternate strategy), across all the tasks. Looking at the exact layers that were dropped: top –

$\{11, 12\}$; even-alternate – $\{10, 12\}$; odd-alternate – $\{9, 11\}$, we can say that (i) dropping last two consecutive layers is more harmful than removing alternate layers, and (ii) keeping the last layer is more important than keeping other top layers. This is contrary to the general understanding that the last layer is well optimized for the pre-trained task and may not be essential for task-specific fine-tuning. At $K = 6$, the *Alternate dropping* strategies show a large drop in the performance, perhaps due to removal of lower layers. Remember our results from bottom strategy showed that lower layers are critical for the model.

The *Symmetric strategy* gives importance to both top and bottom layers and drops the middle layers. Dropping two middle layers from BERT degrades the performance by 0.97 points and makes it the second best strategy at $K = 2$. However, on XLNet the performance degrades drastically when dropping the same set of layers. Comparing these two models, XLNet is sensitive to the dropping of middle layers while BERT shows competitive results to the *Top-layer dropping* strategy even after removing 4 middle layers. We analyze this difference in the behavior of the two models in Section 5.

For *Contribution-based strategy*, we chose layers $\{3, 5\}$ at threshold 0.95 and $\{3, 5, 8, 9\}$ at threshold 0.925 for BERT, and layers $\{9, 10, 11\}$ at threshold 0.925 and $\{8, 9, 10, 11\}$ at threshold 0.9 for XLNet. Using a lower or a higher similarity threshold resulted in dropping none or more than half of the layers in the network. For BERT, the contribution-based dropping did not work well since the method chose a few lower layers for dropping. On the contrary, it worked quite well on XLNet where higher layers were selected. This is in-line with the findings of top and bottom strategy that all models are robust to dropping of higher layers compared to dropping of lower layers.

Due to space limitations, we present the results for RoBERTa, as well as comparison to BERT and DistilRoBERTa in the Appendix. We found that RoBERTa is more robust to pruning layers compared to BERT. Moreover, our 6-layer RoBERTa yielded performance comparable to that of DistilRoBERTa.

DistilBERT: The straight dashed line in Figure 2 compares the result of DistilBERT with the results of pruned versions of 12-layer models. For the same number of layers, a direct comparison can be done between DistilBERT and the yellow bars in

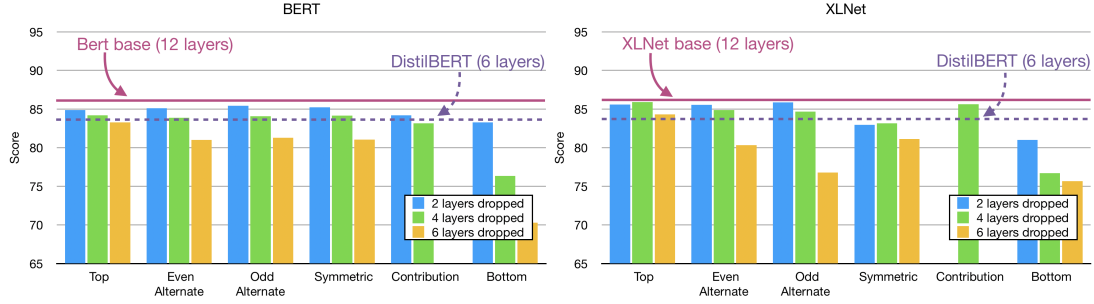


Figure 2: Average classification performance on GLUE tasks when using different layer-dropping strategies and when removing different numbers of layers for BERT and XLNet. Note that the contribution-based strategy selects layers based on the similarity threshold. In some cases it does not select (2,4 or 6) number of layers, which results in some missing bars in the figure.

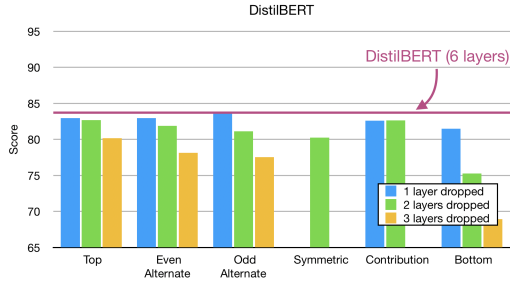


Figure 3: Average classification performance of DistilBERT using different layer-dropping strategies.

the figure, i.e., 6 layers are dropped from the 12-layer networks. It is remarkable to see that our simple *Top dropping strategy yields comparable results to DistilBERT*. Note that building a DistilBERT model consumes a lot of computational power and GPU/TPU memory. In comparison our strategies can be directly applied to the pre-trained models, without the need of any training from scratch.

We further extend our layer-dropping experiments to DistilBERT in order to probe whether distilled models can be safely pruned. Figure 3 presents the average results for DistilBERT. We see similar trends: the top strategy is most consistent, Odd-alternate dropping of single layer improved the performance by 0.22 points over no dropping. DistilBERT serves as an interesting case where a model is designed to be a compressed version of a larger model, but we can still remove one-third of its layers with an average loss of 0.75 points only.

4.2 Task-wise Results

The *Top-layer strategy* works consistently well for all models. In the rest of the paper, we discuss the results for the *Top-layer strategy* only, unless

specified otherwise. Table 1⁵ presents the results for the individual GLUE tasks using the *Top-layer strategy*. We observe the same trend as for the averaged results: *for most of the tasks, we can safely drop half of the top layers in BERT, XLNet or DistilBERT at the cost of losing only 1-3 points*. Note that the QQP and STS-B tasks are least affected by the dropping of layers across all models. When dropping half of the layers from 12-layer models, there was no loss in performance for QQP on XLNet, and only a loss of 0.72 for BERT. Similarly, for STS-B we observe a decrease of only 0.56 and 0.34 points for XLNet and BERT respectively.

We further investigate how many layers are strictly necessary for each individual task. Table 2 shows the results of the minimum number of required layers to maintain 1%, 2% and 3% performance. We found that with XLNet, QQP maintains a performance within 1 point when dropping up to 9 top layers of the model. Essentially, the model consists of only three layers – {1, 2, 3} from the original pre-trained model. We saw a similar trend with BERT, but the drop in performance is in the range of 3 points with a maximum drop of up to 9 top layers. On the other hand, the QNLI and the MNLI tasks are most sensitive to dropping layers across all models. This may reflect that these are more general tasks that require a larger network to perform well. This is inline with recent research that uses MNLI fine-tuned models as a base for other task-specific fine-tuning (Liu et al., 2019c).

To summarize, we found that for 12-layer models, dropping top four layers works consistently well for all tasks with a loss of at most 2%. A task-

⁵We use standard settings provided in the Transformer library to produce the results. This causes a slight mismatch between some of the numbers mentioned in the original papers of each models and our paper.

Drop.	SST-2	MNLI	QNLI	QQP	STS-B	RTE	MRPC
BERT							
0	92.43	84.04	91.12	91.07	88.79	67.87	87.99
2	92.20 (0.23↓)	83.26 (0.78↓)	89.84 (1.28↓)	90.92 (0.15↓)	88.70 (0.09↓)	62.82 (5.05↓)	86.27 (1.72↓)
4	90.60 (1.83↓)	82.51 (1.53↓)	89.68 (1.44↓)	90.63 (0.44↓)	88.64 (0.15↓)	67.87 (0.00)	79.41 (8.58↓)
6	90.25 (2.18↓)	81.13 (2.91↓)	87.63 (3.49↓)	90.35 (0.72↓)	88.45 (0.34↓)	64.98 (2.89↓)	80.15 (7.84↓)
XLNet							
0	93.92	85.97	90.35	90.55	88.01	65.70	88.48
2	93.35 (0.57↓)	85.67 (0.30↓)	89.35 (1.00↓)	90.69 (0.14↑)	87.59 (0.42↓)	66.06 (0.36↑)	86.52 (1.96↓)
4	92.78 (1.14↓)	85.46 (0.51↓)	89.51 (0.84↓)	90.75 (0.20↑)	87.74 (0.27↓)	67.87 (2.17↑)	87.25 (1.23↓)
6	92.20 (1.72↓)	83.48 (2.49↓)	88.03 (2.32↓)	90.62 (0.07↑)	87.45 (0.56↓)	65.70 (0.00)	82.84 (5.64↓)
DistilBERT							
0	90.37	81.78	88.98	90.40	87.14	60.29	85.05
1	90.37 (0.00)	80.41 (1.37↓)	88.50 (0.48↓)	90.33 (0.07↓)	86.21 (0.93↓)	59.93 (0.36↓)	84.80 (0.25↓)
2	90.25 (0.12↓)	79.41 (2.37↓)	86.60 (2.38↓)	90.19 (0.21↓)	86.91 (0.23↓)	62.82 (2.53↑)	82.60 (2.45↓)
3	87.50 (2.87↓)	77.07 (4.71↓)	85.78 (3.20↓)	89.59 (0.81↓)	85.19 (1.95↓)	58.48 (1.81↓)	77.45 (7.60↓)

Table 1: Task-wise performance for the top-layer dropping strategy using the official development sets.

δ	SST-2	MNLI	QNLI	QQP	STS-B
BERT					
1%	2	3	1	7	6
2%	5	4	4	8	7
3%	7	6	5	9	8
XLNet					
1%	5	4	4	9	7
2%	7	5	5	9	8
3%	8	6	7	9	9
DistilBERT					
1%	2	0	1	3	2
2%	2	1	1	3	3
3%	3	2	2	4	3

Table 2: Number of layers dropped from the network while maintaining the performance within a pre-defined threshold δ .

specific optimization of layer-dropping results in a better balance between performance degradation and model size.

4.3 Memory and Speed Comparison

Dropping layers reduces the size of the network and thus reduces the number of parameters, and speeds up the task-specific fine-tuning and the inference time. Table 3 compares the number of parameters, speedup in the fine-tuning process, reduction of inference time, and the loss in performance as the number of layers dropped increases. We can see that **by dropping half of the layers of the network, the average performance drops between 1.81–3.28 points, the number of model parameters reduce by 40% and the fine-tuning and inference time reduces by 50%**. The results for XLNet are remarkable; as with all the memory and speed improvements, the average performance dropped by 1.81 points only.

With a slight increase in memory and computation time at $K = 4$ (dropping a third of the network instead of half), XLNet achieves a performance close to dropping no layers but with 32% reduction in inference time. It is worth reiterating here that a better trade-off between computational efficiency and loss in performance can be achieved by optimizing for a specific task. For example, Table 2 shows that QQP maintains the performance within 1% on XLNet when 9 layers are dropped. This corresponds to 60% reduction in the number of parameters and 80% reduction in terms of inference time.

5 Discussion

Pruned BERT/XLNet vs. DistilBERT Going back to Table 3, we can see that DistilBERT is trained with 6 layers and 66M parameters. After dropping 6 layers from BERT and XLNet, the size of the resulting networks is comparable to DistilBERT. We compile the results for these models in Table 4.⁶ Our six-layer BERT (BERT-6) and XLNet (XLNet-6) showed competitive performance to DistilBERT. They also showed comparable memory and speed performance as shown in Table 3. This result is quite astonishing, given that our reduced models do not require any additional training, while building a DistilBERT model requires training from scratch, which is a time consuming and computation expensive process. Moreover, our

⁶We left out MRPC and RTE of the analysis intentionally, as we found them extremely unstable for all runs including the ones that use the full network. This may be due to the small size of their training data: only 3.6k and 2.5k instances for MRPC and RTE respectively. The results are presented in Table 1.

Drop	Loss	Param.	Fine-tuning speedup	Inference seconds
BERT				
0	0.00	110M	1.00	-
2	1.33	94M	1.24	17% ↓
4	2.00	80M	1.48	33% ↓
6	2.91	66M	1.94	50% ↓
XLNet				
0	0.00	116M	1.00	-
2	0.54	101M	1.20	16% ↓
4	0.23	86M	1.49	32% ↓
6	1.81	71M	1.96	49% ↓
DistilBERT				
0	0.00	66M	1.00	-
1	0.49	59M	1.19	17% ↓
2	0.75	52M	1.48	33% ↓
3	3.28	45M	1.94	50% ↓

Table 3: Comparing the number of parameters (Param.), the speed up in the fine-tuning step, and the inference time for different models. *Fine-tuning speedup* shows the how many times the model speeds up compared to the original network. We report inference time on the QQP devset consisting of 40.4k instances with a batch size of 32.

Model	SST-2	MNLI	QNLI	QQP	STS-B
DistilBERT	90.37	81.78	88.98	90.40	87.14
BERT-6	90.25	81.13	87.63	90.35	88.45
XLNet-6	92.20	83.48	88.03	90.62	87.45

Table 4: Task-wise performance for models of comparable sizes. *-6 refers to models after dropping the top 6 layers.

setup offers the flexibility to choose different sizes of the model based on the computational requirements and the specifics of a downstream task.

BERT vs. XLNet In addition to the reduction benefits of our strategies, they illuminate model-specific peculiarities that help us in comparing and understanding them. Here we discuss our observations and findings about BERT and XLNet. XLNet shows robustness towards dropping top layers (see the relatively smaller drop in performance for XLNet compared to BERT in Table 1). This implies that the **lower layers of XLNet are much richer than those for BERT and are able to learn complex task-specific information** much earlier in the network. We probe this further by building a classifier on individual layers of fine-tuned models and analyzing layer-wise performance. Figure 4 shows the average layer-wise performance of BERT and XLNet. XLNet matures in performance close to the 7th layer of the model while BERT improves with ev-

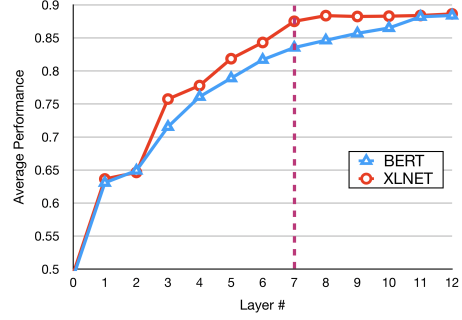


Figure 4: Average layer-wise classification results.

ery higher layer until the 11th layer. This suggests that (i) XLNet learns task-specific knowledge at much lower layers compared to BERT, (ii) the last layers of XLNet might be quite redundant and are a good candidate for dropping without a large drop in performance. The second point is well in line with our contribution-based strategy where we picked the top layers of XLNet for removal. Remember that the *Contribution-based strategy* is motivated by how much value a layer is adding to the network, which was estimated by measuring the similarity between the input and the output representations of that layer.

Recent studies on BERT have shown that during fine-tuning most of the changes happen at the higher layers with marginal changes in the lower layers (Kovaleva et al., 2019; Houlsby et al., 2019). It is unclear whether this also holds for XLNet, as we observed that the model matured close to the middle layers. To shed light on this, we calculated the layer-wise cosine similarity between the pre-trained model and its fine-tuned version for both BERT and XLNet. Figure 5 shows the similarity curve averaged across all tasks. As expected, **the similarity between base and fine-tuned BERT models reduces from lower to higher layers** and is aligned with the findings of Kovaleva et al. (2019) and others. However, XLNet shows a completely different behavior. The middle layers undergo major changes while the lower and higher layers remain relatively close to the base model. This explains the large degradation in performance for XLNet when only two middle layers were dropped as can be seen in Figure 2 with symmetric dropping at $K = 2$.

From a theoretical point of view, XLNet and BERT are quite different in design, even though both use a Transformer-based architecture. XLNet is an auto-regressive model that uses permutation operations to learn context, while BERT is an auto-

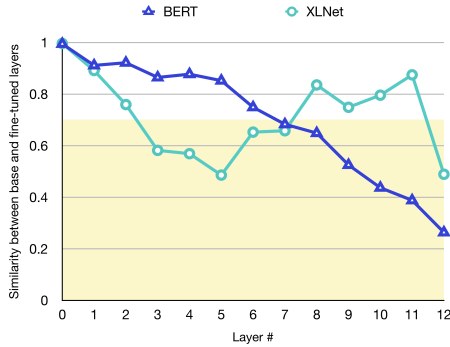


Figure 5: Layer-wise average cosine similarity between pre-trained and fine-tuned models. The yellow shaded region highlights layers that changed substantially after fine-tuning (less than 0.7 cosine similarity).

encoder that relies on predicting a masked token to do the same. We speculate that the presence of explicit contextual information, via all possible permutations of the factorization order in the training process enables XLNet to learn sequence information at lower layers of the network. This could potentially be one of the reason that purely attention-based models such as BERT require very deep architectures in order to learn the contextual information, and why LSTM-based models such as ELMo (Peters et al., 2018) do not.

Layer-Dropping using Fine-tuned Models

One of the advantages of our dropping strategies is that they are directly applied to the pre-trained models, i.e., we avoid the need to optimize our strategies for each task. However, it is possible that dropping from a fine-tuned model may result in better performance. To explore this idea, we tried our dropping strategies on fine-tuned models. More specifically, **we first fine-tune the model, drop the layers, and then fine-tune the model again**. Table 5 presents this method on BERT and XLNet. We found this setup to be comparable to dropping layers directly from the pre-trained model in most of the cases. The result also shows that our method of dropping layers directly from a pre-trained model does not lose any critical information which was essential for a specific task. However, **we do think that pruning a fine-tuned model may lose task-specific information since after fine-tuning, the model is optimized for the task. This is reflected in some of the results of BERT/XLNet-FT-6**. Other disadvantages of this method are that (i) **it builds task-specific reduced models instead of one general reduced model** that

Model	SST-2	MNLI	QNLI	QQP	STS-B
BERT-6	92.25	81.13	87.63	90.35	88.45
BERT-FT-6	90.02	80.85	87.24	90.34	88.16
XLNet-6	92.20	83.48	88.03	90.62	87.45
XLNet-FT-6	92.43	83.75	86.80	90.77	87.60

Table 5: Task-specific dropping. XLNet-FT-6 first fine-tunes the pre-trained model by freezing the layers to drop, and then it removes the layers and performs fine-tuning again.

can be used for task-specific fine-tuning, and (ii) it requires running **fine-tuning twice for each task**, which is a time-consuming process.

Iterative Dropping Similarly to task-specific dropping, here we want to preserve the model’s performance during the dropping process. Instead of dropping all layers together, we iteratively drop one layer after every two epochs of the fine-tuning process. This did not yield any improvements over dropping layers directly from the pre-trained model.

6 Related Work

Along with a rapid influx of pre-trained neural language models, there has also been a surge towards down-scaling these models to adapt them in computationally limited configurations such as hand-held devices. To this end people have explored various methods (Ganesh et al., 2020) to achieve this goal. A popular framework, based on knowledge distillation also known as student-teacher model (Hinton et al., 2015) trains a student model to replicate the distribution of the large-scale teacher model. Several researchers applied task-agnostic distillation to the pre-trained BERT model (Sanh et al., 2019; Sun et al., 2019, 2020). Others attempted to distill task-specific BERT for NLU tasks (Jiao et al., 2019), refined by multi-task learning on the GLUE benchmark (Liu et al., 2019a). Several researchers have achieved compression through quantization (Zafir et al., 2019; Shen et al., 2019).

Another line of work aims at reduction by pruning unimportant parts of the network. Researchers have demonstrated ablating attention-heads (Michel et al., 2019a; Voita et al., 2019a,b), unimportant neurons (Bau et al., 2019; Dalvi et al., 2019). Gordon et al. (2019) performed applied *Magnitude Weight Pruning* by pruning weights below a certain threshold. Fan et al. (2019) introduced *LayerDrop* during training and showed that it en-

courages robustness in the model and facilitates dropping of layers at inference time with minimal impact on performance. In contrast, we perform layer dropping on the trained model, which both yields smaller models (their model is the same size as BERT) and also does not require pre-training from scratch. Compared to their 6-layer RoBERTa model, our 6-layer RoBERTa system resulted in 1.54 points better performance on the MNLI task and is only 0.53 points lower on the SST-2 task. Finally Lan et al. (2019) proposed a lighter BERT by factorizing weight matrix for the embedding layer into two smaller matrices and sharing weight matrices across hidden layers of the model.

Our work falls in the class of pruning methods. The advantage of our proposed pruning strategies is that they do not require training a model from scratch. Compared to task-specific pruning, our setup does not require to build separate fine-tuned pruned models and works directly on the pre-trained models. Our setup is complementary to other distillation methods such as quantization and student-teacher learning and thus it can be combined with them.

7 Conclusion

We proposed layer pruning strategies that do not require training a new model, but can be directly applied to pre-trained models. Our best pruning strategy (top-layer dropping) achieved a 40% reduction in model size and 50% reduction in inference time, while maintaining up to 98.2% of the original accuracy across well known architectures, BERT, XLNet and RoBERTa. Our reduced BERT model and RoBERTa model achieved comparable results to that of DistilBERT and DistilRoBERTa respectively, in terms of GLUE performance, time for fine-tuning and inference time. However, unlike DistilBERT and DistilRoBERTa, our method does not require re-training from scratch, which is time-consuming and memory-intensive. Our approach also offers a trade-off between accuracy and model size.

In addition, we provided a detailed comparison between BERT and XLNet, illuminating insightful understanding of these models. We show that XLNet is much more robust to pruning because it learns sequence-level knowledge earlier in the network compared to BERT.

Our findings encourage new smarter ways to apply student-teacher architecture. While previous

work, e.g., Sun et al. (2019), made the student initialized by taking one layer out of every two layers. Here we have presented evidence for the need to take all lower layers into account. In future work, we plan to experiment with other NLP tasks and other Transformer architectures, as well as with combining layer dropping with attention head removal.

References

- Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. [Identifying and controlling important neurons in neural machine translation](#). In *International Conference on Learning Representations*.
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. [What do Neural Machine Translation Models Learn about Morphology?](#) In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver. Association for Computational Linguistics.
- Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth pascal recognizing textual entailment challenge. In *In Proc Text Analysis Conference (TAC09)*.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. [SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.
- Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, D. Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI, Oral presentation)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.
- William B. Dolan and Chris Brockett. 2005. [Automatically constructing a corpus of sentential paraphrases](#). In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
- Angela Fan, Edouard Grave, and Armand Joulin. 2019. [Reducing transformer depth on demand with structured dropout](#).

- Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Haris Ali Khan, Yin Yang, Deming Chen, Marianne Winslett, Hassan Sajjad, and Preslav Nakov. 2020. Compressing large-scale transformer-based models: A case study on bert. *ArXiv*, abs/2002.11985.
- Mitchell A. Gordon, Kevin Duh, and Nicholas Andrews. 2019. Compressing BERT: Studying the effects of weight pruning on transfer learning. *ArXiv*, abs/2002.08307.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). Cite arxiv:1503.02531Comment: NIPS 2014 Deep Learning Workshop.
- 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. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799, Long Beach, California, USA. PMLR.
- 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](#).
- Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. [Revealing the dark secrets of BERT](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4364–4373, Hong Kong, China. Association for Computational Linguistics.
- 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](#).
- Linqing Liu, Huan Wang, Jimmy Lin, Richard Socher, and Caiming Xiong. 2019a. [Attentive student meets multi-task teacher: Improved knowledge distillation for pretrained models](#).
- Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019b. [Linguistic knowledge and transferability of contextual representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- J. S. McCarley. 2019. [Pruning a bert-based question answering model](#).
- Paul Michel, Omer Levy, and Graham Neubig. 2019a. [Are sixteen heads really better than one?](#) In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14014–14024. Curran Associates, Inc.
- Paul Michel, Omer Levy, and Graham Neubig. 2019b. [Are sixteen heads really better than one?](#) *CoRR*, abs/1905.10650.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#).
- 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](#).
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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, Seattle, Washington, USA. Association for Computational Linguistics.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. [Patient knowledge distillation for BERT model compression](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4322–4331, Hong Kong, China. Association for Computational Linguistics.
- Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. [mobilebert: Task-agnostic compression of bert by progressive knowledge transfer](#). In *International Conference on Learning Representations*.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. [Distilling task-specific knowledge from BERT into simple neural networks](#). *CoRR*, abs/1903.12136.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China.

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](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#).

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. [Q8bert: Quantized 8bit bert](#).

with RoBERTa, the better optimizer introduced in RoBERTa makes it more stable and less vulnerable with respect to layer dropping. If we look at the five stable tasks (excluding RTE and MRPC), the performance of the RoBERTa model dropped by up to 1.73 points absolute, while reducing the number of layers to half. Note that RoBERTa with 6 layers dropped (i.e., RoBERTa-6) is comparable to DistilRoBERTa in terms of performance, number of layers, and parameters. It is worth noting that we did not optimize the parameter values while applying layer-dropping strategies, using the default setting in the Transformer library.

A Appendices

Table 6 summarizes the results of applying the top layer-dropping strategy on the BERT and the RoBERTa models. We added official DistilRoBERTa results⁷ in order to highlight the effectiveness of our strategy. Comparing BERT

⁷<https://github.com/huggingface/transformers/tree/master/examples/distillation>

Drop.	SST-2	MNLI	QNLI	QQP	STS-B	RTE	MRPC
BERT							
0	92.43	84.04	91.12	91.07	88.79	67.87	87.99
2	92.20 (0.23↓)	83.26 (0.78↓)	89.84 (1.28↓)	90.92 (0.15↓)	88.70 (0.09↓)	62.82 (5.05↓)	86.27 (1.72↓)
4	90.60 (1.83↓)	82.51 (1.53↓)	89.68 (1.44↓)	90.63 (0.44↓)	88.64 (0.15↓)	67.87 (0.00)	79.41 (8.58↓)
6	90.25 (2.18↓)	81.13 (2.91↓)	87.63 (3.49↓)	90.35 (0.72↓)	88.45 (0.34↓)	64.98 (2.89↓)	80.15 (7.84↓)
RoBERTa							
0	92.20	86.44	91.73	90.48	89.87	68.95	88.48
2	93.46 (1.26↑)	86.53 (0.09↑)	91.23 (0.50↓)	91.02 (0.54↑)	90.21 (0.34↓)	71.84 (2.89↑)	89.71 (1.23↑)
4	93.00 (0.80↑)	86.20 (0.24↓)	90.57 (1.16↓)	91.12 (0.64↑)	89.77 (0.10↓)	70.40 (1.45↑)	87.50 (0.98↓)
6	91.97 (0.23↓)	84.44 (2.00↓)	90.00 (1.73↓)	90.91 (0.43↑)	88.92 (0.95↓)	64.62 (4.33↓)	85.78 (2.70↓)
DistilRoBERTa							
0	92.5	84.00	90.80	89.40	88.30	67.90	86.60

Table 6: Comparing the task-wise performance of BERT, RoBERTa for the top-layer dropping strategy with DistilRoBERTa using the official development sets.