On the Importance of Data Size in Probing Fine-tuned Models

Anonymous ACL submission

Abstract

Several studies have investigated the reasons behind the effectiveness of fine-tuning, usually through the lens of probing. However, these studies often neglect the role of the size of the dataset on which the model is fine-tuned. In this paper, we highlight the importance of this factor and its undeniable role in probing performance. We show that the extent of encoded linguistic knowledge depends on the number of fine-tuning samples, specifically the number of iterations for which the model is updated. The analysis also reveals that larger training data mainly affects higher layers, and that the extent of this change is a factor of the number of iterations in fine-tuning rather than the diversity of the training samples. Finally, we show through a set of experiments that fine-tuning introduces shallow and recoverable changes to model's representation.

1 Introduction

The outstanding performance of pre-trained language models (LMs) on many NLP benchmarks has provoked curiosity about the reasons behind their effectiveness. To this end, several probes have been proposed to explore their capacity (Tenney et al., 2019b; Hewitt and Manning, 2019; Wu et al., 2020). The investigations have clearly highlighted the abilities of LMs in capturing various types of linguistic knowledge (Liu et al., 2019; Clark et al., 2019; Michael et al., 2020; Klafka and Ettinger, 2020; Tenney et al., 2019a).

However, to take full advantage of the encoded knowledge of pre-trained models in specific target tasks, it is usually required to perform a further fine-tuning (Devlin et al., 2019). The broad application of fine-tuning has garnered the attention of many researchers to explore its peculiarities. Trying to understand the fine-tuning procedure, recent analyses have shown that most of the pre-trained linguistic knowledge is preserved after fine-tuning (Tenney et al., 2019b). Furthermore, by encoding the essential linguistic knowledge in the lower layers, this procedure makes the upper layers task-specific (Durrani et al., 2021). Focusing on the role of the encoded knowledge in the probing accuracy, Mosbach et al. (2020) introduce the attention distribution as an effective factor on probing performance of fine-tuned models.

In this work, we present another important factor in interpreting probing results for fine-tuned models. Our investigations reveal that the conclusions drawn by previous probing studies that investigate the impact of fine-tuning on acquiring or forgetting knowledge might not be fully reliable, unless the size of the fine-tuning dataset is also taken into account. Through several experiments, we show that the encoded linguistic knowledge can highly depend on the size of target tasks' datasets. Specifically, the larger the task data, the more the probing performance deviates from the pre-trained model, irrespective of the fine-tuning tasks.

To address the overlooked role of data size, we run several experiments by limiting training samples and probing the fine-tuned models. Our results indicate that models fine-tuned on large training datasets witness more change in their linguistic knowledge compared to pre-trained BERT. However, by reducing fine-tuning training data size (e.g., from 393k in MNLI to 7k), the gap between probing scores becomes smaller. Moreover, we expand our analysis and evaluate to what extent large training datasets affect the captured knowledge across layers. The layer-wise results show that the effect of data size is more notable on higher layers. Also, this pattern is significantly obvious in the models fine-tuned by a larger dataset. We take our analysis a step further, and show that the difference in probing performance among different data sizes are due to the total number of optimization steps rather than the diversity of training samples. However, we have realized that the modifications from fine-tuning is somehow shallow, to the extent that the linguistic knowledge can be recovered even after fine-tuned on several tasks.

The findings of this paper can be summarized as follows:

- Data size is a factor that highly impacts fine-tuned model's linguistic knowledge.
- Higher layers are the most susceptible layers to data size.
- The number of training steps are actually what makes larger datasets have higher impacts on the model's linguistic knowledge (rather than the diversity in training samples).
- The linguistic knowledge introduced to a model by a fine-tuning task can be retrieved through re-fine-tuning even after sequentially fine-tuning on other downstream tasks.

2 Related Work

Recently, many studies have shown that pre-trained language models, such as BERT (Devlin et al., 2019), encode certain linguistic knowledge in their internal representations (Tenney et al., 2019b). For instance, Hewitt and Manning (2019) found that syntactic dependencies can be obtained from BERT's token embeddings, suggesting that BERT encodes syntactic knowledge in its representations. Nevertheless, not all layers behave similarly in capturing linguistic features: lower layers tend to encode surface-level knowledge, middle layers seem to be responsible for syntactic information and higher layers capture semantic knowledge in their representations (Jawahar et al., 2019).

While models such as BERT capture considerable amount of linguistic features, one still requires to fine-tune them to take advantage of their full potential in specific downstream tasks (Wang et al., 2018). Fine-tuning affects BERT in various ways, for instance, Hao et al. (2020) found that fine-tuning mainly affects the attention mode of last layers and altering the feature extraction mode of the middle and last layers. In addition, fine-tuning BERT on a negation scope task improves the model's attention sensitivity to negation (Zhao and Bethard, 2020).

Apart from the changes made to BERT's attention, recent work has studied how fine-tuning affects BERT's representations and, as a result, its linguistic knowledge. Merchant et al. (2020) found that fine-tuning primarily affects the representations in higher layers, and depending on the downstream task, the changes made to lower layers could be either deep or shallow. Moreover, on only a small number of downstream tasks, fine-tuning seems to have a positive impact on the probing accuracy (Mosbach et al., 2020). Given the fact that fine-tuning mostly affects higher layers, Durrani et al. (2021) showed that after fine-tuning most of the model's linguistic knowledge is transferred to lower layers to reserve the capacity in the higher layers for task-specific knowledge.

	Full	**7k**	**2.5k**	**1k**
CoLA	57.55	56.87	46.68	42.72
SST-2	92.78	91.28	89.79	86.81
MNLI	83.19	73.73	68.63	60.16
QQP	90.63	82.37	79.93	76.93
MRPC	86.43	-	81.78	77.82

Table 1: The performance of fine-tuned BERT on five tasks from GLUE (dev set) after fine-tuning on training data of varying size. The numbers are reported based on accuracy for SST, MNLI, QQP, MRPC and Matthew's correlation for CoLA.

Studies so far have relied on probing accuracy to explain how fine-tuning affects a model's linguistic knowledge (Mosbach et al., 2020; Durrani et al., 2021; Merchant et al., 2020). However, given the fact that fine-tuning tasks do not share the same number of samples, concluding to what extent target tasks contribute to the model's linguistic knowledge is not fully reliable. To the best of our knowledge, none of the previous studies have considered the role of data size in fine-tuned models' linguistic knowledge. In this work, we show that the size of the dataset plays a crucial role in the amount of knowledge captured during fine-tuning. By designing different experiments, we analyze the effect of the size of the dataset in-depth.

3 Experimental Setup

We have carried out over 600 experiments to study the linguistic features captured during fine-tuning. This allows us to examine how much different factors impact performance on different probing tasks. Moreover, varying the sample size lets us understand its importance in analyzing fine-tuned models. In this section, we provide more details on setups, downstream tasks, and probing tasks.

3.1 Fine-tuning

For our analyses, we concentrate on the BERT-base model, which is arguably the most popular pre-trained model. We fine-tuned the 12-layer BERT on a set of tasks from the GLUE Benchmark (Wang et al., 2018) for five epochs and saved the best checkpoint based on performance on the validation set. We used the [CLS] token for classification and set the learning rate as $2e^{-10}$. We have chosen the following target tasks:

CoLA. The Corpus of Linguistic Acceptability is a binary classification task in which **8.5k** training samples are labeled based on their grammatical correctness (Warstadt et al., 2019).

MRPC. The Microsoft Research Paraphrase Corpus includes **3.6k** training sentence pairs in which the semantic equivalence of two sentences is determined (Dolan and Brockett, 2005).

SST-2. The Stanford Sentiment Treebank is a sentiment classification task containing **67k** training sentences (Socher et al., 2013).

QQP. With **364k** question pairs, the goal of the Quora Question Pairs dataset is to determine whether two questions in a pair are semantically similar.

MNLI. The Multi-Genre Natural Language Inference is a Natural Language Inference (NLI) task with about **393k** records in its training set (Williams et al., 2018).

3.2 Fine-tuning performance

The performance of the fine-tuned models on these tasks is illustrated in Table 1. We report the results on different training data sizes[1] to highlight the extent to which reducing training data affects a model's performance on the corresponding tasks. It is worth mentioning that even though the performance of target tasks decreases by reducing their training data, it is still far better than the pre-trained version. Therefore, the models have learned the corresponding target tasks to some extent.

3.3 Probing tasks

We probe the pre-trained and fine-tuned BERT models by training a linear classifier on top while the weights of the encoders are frozen. Keeping the probing classifier simple lets us scrutinize the linguistic knowledge by eliminating the possibility of the classifier learning itself. All probes are trained with a batch size of 32, a learning rate of $3e^{-4}$, for 10 epochs. Due to limited resources, we fine-tuned models with three random seeds and probed selected ones with three random seeds to determine the noise in probing accuracy. The probing accuracy remained stable, ranging within ± 1.0. Finally, we report the evaluation scores on test sets for the models with the highest validation accuracy on the validation set.

We opted for four syntactic and semantic probing tasks from the SentEval benchmark (Conneau and Kiela, 2018) to study the linguistic knowledge encoded in the models. The binary classification tasks are as follows:

Bigram Shift is a task that aims to test the model's ability to predict whether two random successive tokens in the same sentence have been inverted.

Object Number focuses on the model's ability to determine the singularity or plurality of the main clause's direct object.

Coordination Inversion examines the model's ability to distinguish between original sentences and sentences where the order of two coordinated clausal conjoints have been inverted.

Semantic Odd Man Out is a task that tests the model's ability to predict if a sentence is original or whether a random word has been replaced with another word from the same part of speech.

4 Data Size Analysis

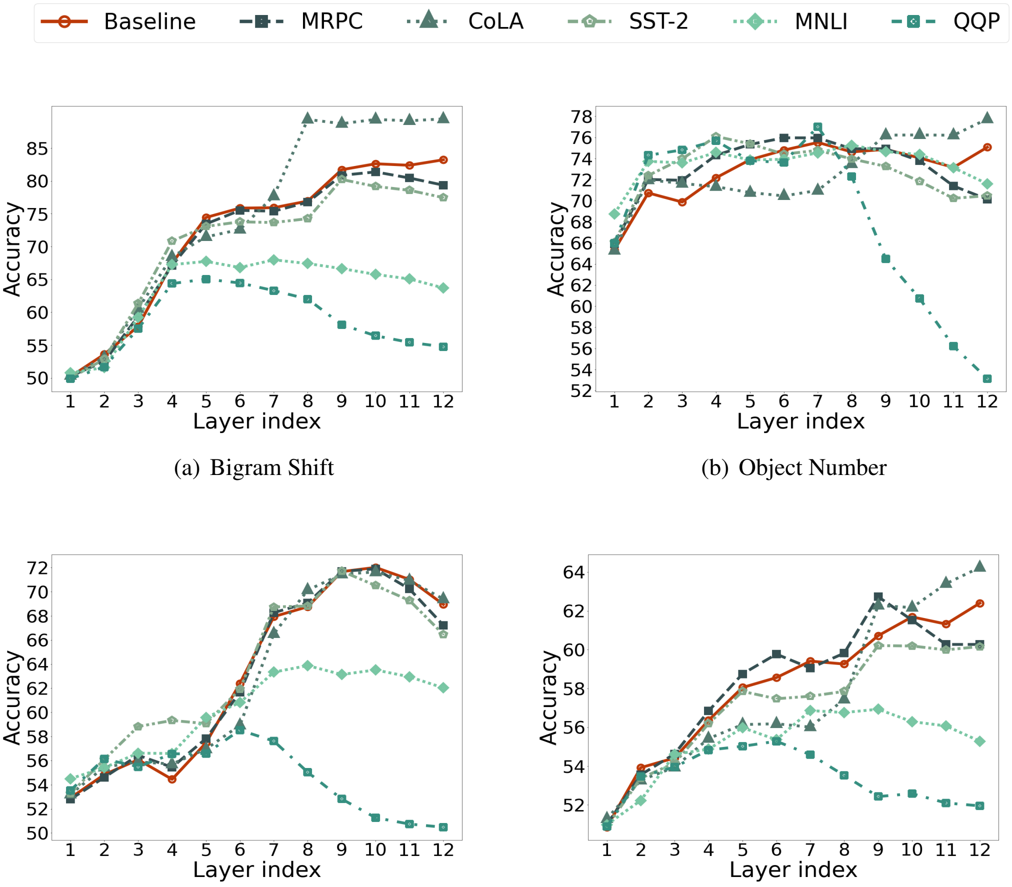
In this section, we first provide insight on the role of target tasks in capturing or forgetting different types of knowledge (e.g., syntactic and semantic) during fine-tuning. Then, we investigate the role of datasets' sizes on linguistic knowledge.

4.1 Probing Linguistic Knowledge

We empirically evaluate the linguistic knowledge captured by several fine-tuned models through the means of probing.

Figure 1 illustrates the layer-wise probing performance of fine-tuned models, considering pre-trained BERT as our baseline. As can be observed, different models carry similar linguistic knowledge

[1] Since MRPC only has 3.6k training samples, we do not report any 7k results for this dataset.



(a) Bigram Shift

(b) Object Number

(c) Coordination Inversion

(d) Semantic Odd Man Out

Figure 1: Probing accuracy on all the layers of fine-tuned models. As shown, there is a large accuracy gap between models fine-tuned on larger data sizes (e.g., MNLI and QQP) and the baseline.

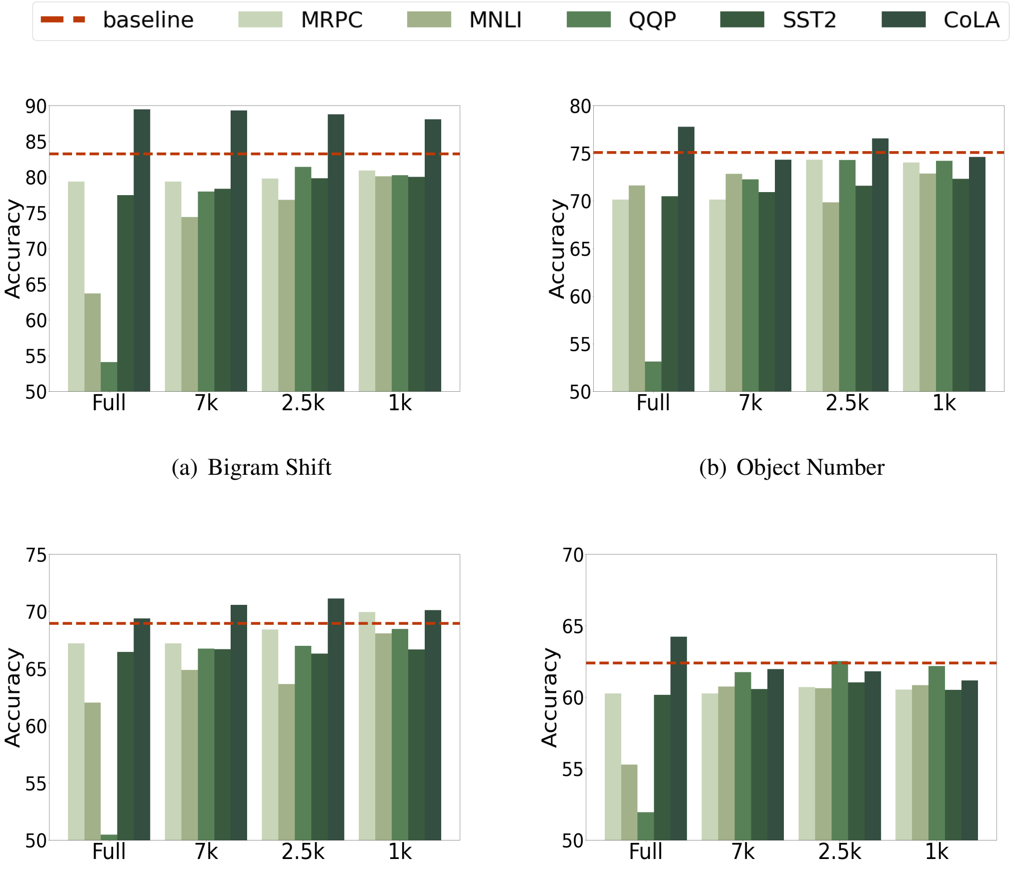
up to the middle layers, and the difference gradually increases as we move up to the higher layers. This observation is consistent with the reported results by Merchant et al. (2020). Their experimental analysis indicates that fine-tuning mostly changes the higher layers while having very less impact on the lower layers. Durrani et al. (2021) also reported a similar behavior in other LMs through different probing tasks.

The results illustrated in Figure 1 clearly show that how data size impacts probing accuracy. As stated in Section 3.1, fine-tuning tasks contain different number of samples, some of which are much larger than the others. We can witness that the probing performance of the baseline and models fine-tuned on small datasets are within a close range, while fine-tuning on larger data sizes (e.g., QQP and MNLI) can significantly impact the models' linguistic knowledge. Following this interesting pattern, we carry out a set of experiments to understand whether the mentioned pattern in the models' linguistic knowledge can be due to different data sizes.

4.2 The Impact of Data Size

One of the popular studies in probing is to check fine-tuned models for specific linguistic knowledge. The changes brought to the model upon fine-tuning are taken as a means to explain the nature of the corresponding task on which fine-tuning has been done. Existing studies usually consider several tasks, many of which do not have datasets of comparable size. For instance, MNLI is 46 times larger than CoLA. Regardless of the number of samples that every target task has, previous studies have only relied on the type of downstream tasks. Therefore, they can not answer why some target tasks cause more profound modifications to the encoded linguistic knowledge compared to others.

The results of Section 4.1 reinforce the hypothesis that the number of samples (data size) could be an important cause of improving or impairing the



(a) Bigram Shift

(b) Object Number

(c) Coordination Inversion

(d) Semantic Odd Man Out

Figure 2: An illustration of models' performance fine-tuned on fixed-size training set across four probing tasks. The pre-trained BERT's performance has been shown by dashed red line. The figures suggest that different fine-tuned models almost encode similar linguistic knowledge, specifically semantic knowledge, when they are trained with equal size training data.

linguistic knowledge captured during fine-tuning. In Figure 1, we observe that there is a significant difference between the probing accuracy of models fine-tuned on large datasets (e.g., MNLI and QQP) and the ones on small datasets (e.g., MRPC). So, our hypothesis can explain this significant difference in the encoded linguistic knowledge across fine-tuned models.

We examine our hypothesis by fine-tuning pre-trained BERT on the selected downstream tasks with a set of different number of samples. Taking the pre-trained BERT as the baseline, we analyze the effect of the training set size on the encoded linguistic knowledge by limiting the number of samples to 7k, 2.5k, and 1k. Figure 2 is an illustration of our experiments regarding the data size's effect on the encoded linguistic knowledge. These results confirm our hypothesis that data size in fact plays a significant role in probing accuracy. We further elaborate on this effect in the following discussion.

4.3 Discussion

The effect of data size on both the syntactic and semantic probing tasks is notable. However, the difference is more significant on syntactic knowledge, Figure 2(a). This could be attributed to the model's resistance to losing its semantic knowledge, as witnessed by a more stable performance in semantic probing tasks, in Figures 2(b), 2(c), and 2(d).

We observe that as the number of samples increases, the gap between fine-tuned models and the pre-trained BERT (baseline) becomes more apparent. For instance, probing the model fine-tuned on QQP's full training set demonstrates that it has far less linguistic knowledge than the baseline. However, after fine-tuning the model on QQP with fewer training samples (7k, 2.5, and 1k), the results assimilate to each other. This shows that fine-tuning data size indeed affects the linguistic knowledge

		Bigram Shift					Semantic Odd Man Out				
		Full	**7k**	**2.5k**	**1k**	**baseline**	**Full**	**7k**	**2.5k**	**1k**	**baseline**
CoLA	Layer 2	-0.49	0.16	-0.63	-0.82	53.60	-0.65	-0.25	-0.06	-0.23	53.92
	Layer 7	1.78	1.36	1.57	2.03	75.93	-3.40	-2.31	-0.80	-1.43	59.41
	Layer 11	6.78	7.09	6.29	5.10	82.39	2.08	1.78	1.83	0.98	61.32
	Layer 12	6.22	6.09	5.56	4.85	83.23	1.84	-0.44	-0.58	-1.23	62.40
SST2	Layer 2	-0.74	-0.82	-0.30	-0.94	53.60	-0.55	-0.55	-0.52	-0.10	53.92
	Layer 7	-2.26	-1.94	-1.94	-0.24	75.93	-1.81	-1.56	-1.29	-1.22	59.41
	Layer 11	-3.81	-2.48	-1.89	-1.33	82.39	-1.33	-0.87	-0.88	-0.55	61.32
	Layer 12	-5.77	-4.87	-3.40	-3.20	83.23	-2.24	-1.83	-1.37	-1.89	62.40
MNLI	Layer 2	-2.01	-0.78	-0.32	0.51	53.60	-1.69	-0.38	-0.62	-0.13	53.92
	Layer 7	-7.94	-1.68	-0.85	-0.83	75.93	-2.55	-0.54	-0.74	-2.61	59.41
	Layer 11	-17.31	-6.54	-4.49	-1.52	82.39	-5.25	-0.32	-1.30	-0.45	61.32
	Layer 12	-19.52	-8.84	-6.44	-3.14	83.23	-7.12	-1.65	-1.76	-1.55	62.40
QQP	Layer 2	1.93	0.68	0.35	-0.26	53.60	-0.46	-0.12	-0.27	-0.21	53.92
	Layer 7	-12.63	-1.55	-0.05	0.60	75.93	-4.82	-0.01	0.30	-0.53	59.41
	Layer 11	-26.97	-3.78	-1.05	-2.46	82.39	-9.22	0.89	0.90	0.65	61.32
	Layer 12	-29.12	-5.70	-1.81	-3.00	83.23	-10.45	-0.65	0.13	-0.22	62.40
MRPC	Layer 2	-1.08	—	-0.82	-0.96	53.60	-0.37	—	-0.56	-0.53	53.92
	Layer 7	-0.53	—	-1.04	-0.09	75.93	-0.36	—	0.29	-0.34	59.41
	Layer 11	-1.94	—	-1.9	-1.41	82.39	-1.05	—	1.36	1.35	61.32
	Layer 12	-3.87	—	-3.45	-2.31	83.23	-2.13	—	-1.7	-1.86	62.40

Table 2: Layer-wise performance of models on the probing tasks. Each cell represents the difference (delta) in performance between the corresponding fine-tuned model and the baseline. The pre-trained BERT performance (baseline) is shown in the right columns.

encoded by the model.

Overall, in this section, we have uncovered the role of data size in affecting the amount of linguistic knowledge through fine-tuning. This suggests that data size should be taken into account when analyzing fine-tuned models. We will study this effect by individually probing each layer through further experiments.

5 Layer-wise Analysis

Given the previous observations (Figure 2) that data size affects the linguistic knowledge captured by BERT through fine-tuning, we would like to see on which layers these changes are more significant.

As Jawahar et al. (2019) stated, BERT layers are divided into three classes in terms of the linguistic knowledge they capture. To this end, we probe layer 2 (lower layers), layer 7 (middle layers), and layers 11, 12 (higher layers) to demonstrate the changes that data size applies to each category of layers.

Table 2 depicts our results obtained from this experiment, which are compared with BERT-base. Due to our limited resources and an excessive number of experiments, we discard probing tasks that have no distinguished patterns in the previous sections (Figure 1 and 2). Hence, we have omitted Coordination Inversion and Object Number from the probing tasks.

The heatmap follows a similar trend to the one depicted in Figure 2. As we decrease the number of training samples, the probing performance on the fine-tuned models becomes closer to the baseline across all layers. MNLI and QQP's behavior are compelling evidence of the effectiveness of data size across layers. Such models fine-tuned with larger datasets undergo more considerable changes than those with smaller data sizes.

Regardless of data size, we can also observe that fine-tuning mostly affects higher layers. Our finding is aligned with Merchant et al. (2020) that fine-tuning has a greater impact on higher layers and negligible effects on lower layers.

There is also an interesting pattern concerning CoLA's performance. Though its performance drops for about 15 scores from the full to 1k version (Table 1) its linguistic knowledge has been negligibly affected by data size. We leave the investigation on CoLA's interesting behavior to future work.

6 Fixed Iteration Analysis

Given the observations from Section 5, we have realized that by training BERT on larger datasets, the model's performance deviates substantially from the baseline. However, by reducing the size of train-

		Full	7k	2.5k
Bigram Shift				
QQP	Layer 2	52.87	0.07	-0.03
	Layer 7	71.88	-2.08	-1.12
	Layer 11	74.08	0.49	2.90
	Layer 12	73.25	-0.10	1.81
MNLI	Layer 2	51.9	-0.24	-1.16
	Layer 7	71.03	0.88	-0.02
	Layer 11	67.69	1.93	2.47
	Layer 12	65.82	1.48	1.57
Semantic Odd Man Out				
QQP	Layer 2	53.73	0.73	0.49
	Layer 7	56.12	0.95	1.61
	Layer 11	58.11	1.23	1.16
	Layer 12	58.03	1.34	0.31
MNLI	Layer 2	53.23	0.24	0.76
	Layer 7	57.00	1.54	1.60
	Layer 11	57.27	2.10	1.17
	Layer 12	56.77	2.43	1.22

Table 3: The performance of models trained with fixed and equal number of iterations across different sizes on each downstream task. Every cell demonstrates the difference (delta) between the full and the fixed-sized models. With an equal number of iterations, in each layer, fine-tuned models have a similar performance.

ing data, the gap between fine-tuned model and the baseline decreases. This behavior could be due to either the diversity of training samples or the larger number of iterations through which the model is updated.

To factor out the role of the number of iterations, we repeat the same experiment carried out in Section 5 by fixing the number of iterations on all data sizes. This will allow the model to be updated equally across different data sizes within a task. Consequently, this experiment will determine which of the mentioned hypotheses best explains the large gap between the baseline and the full models. Note that we fine-tuned the full models for just 1 epoch to avoid a large number of iterations for the 7k and 2.5k models.

Since SST2, CoLA, and MRPC have notably smaller datasets, and the number of iterations does not differ across the full, 7k, and 2.5k models, we have dropped them from this scenario.

Table 3 summarizes our results. The first interesting pattern is that fine-tuning for more epochs impairs the captured linguistic knowledge significantly. As an instance, we can observe the impact of longer training by comparing Bigram Shift performance in the last layer on the full version of QQP in Table 2 (54.11) and Table 3 (73.25).[2]

As Table 3 suggests, fixing the number of iterations reduces the gap across different data sizes, causing the 7k and 2.5k models to behave almost similarly to the full models. For instance, in Table 2, there is about a difference of 24 scores in the last layer's performance between the full and the 7k QQP on Bigram Shift, which has been reduced to approximately -0.1 with equal training steps, Table 2.

This finding is interesting because, firstly, it indicates that the high variance between baselines and full models is mainly due to the number of times their weights are updated during fine-tuning rather than the diversity of the training samples. Secondly, with equal data sizes, the role of target tasks becomes less influential in the linguistic knowledge introduced into the model by fine-tuning, reinforcing the conclusions from Section 5.

7 Sequence Analysis

Our previous results indicate that the size of fine-tuning data indeed affects the encoded linguistic knowledge in the higher layers of pre-trained BERT. In this section, we investigate whether data size has the same effects on re-fine-tuning a model as it has on fine-tuning a pre-trained model for the first time and whether these changes can be recovered.

To address this question, we have designed an experiment in which we fine-tune BERT sequentially on CoLA and SST2, and once again on CoLA (CoLA → SST2 → CoLA). We also carried out the same procedure with SST2 → CoLA → SST2. By probing the final models, we verify the role played by data size on manipulating model's knowledge captured during fine-tuning.

Results are reported in Table 4. We can see that whenever we re-fine-tune a model, its linguistic knowledge is replaced by the latest fine-tuning task. For example, Bigram Shift accuracy of BERT fine-tuned on CoLA is 89.45, but with re-fine-tuning on SST-2, the accuracy drops to 79.59, which is almost similar to SST-2's accuracy on Bigram Shift. This means that the knowledge introduced to the

[2] As mentioned in Section 3.1, the models in Table 2 were fine-tuned for five epochs.

	BShift	ObjNum	CoordInv	SOMO
SST-2	77.46	70.48	66.46	60.16
CoLA	89.45	77.76	69.40	64.24
SST2 → CoLA	87.96	73.68	63.19	63.19
CoLA → SST2	79.59	72.21	65.91	61.58
SST2 → CoLA → SST2	79.11	70.37	66.10	60.06
CoLA → SST2 → CoLA	88.13	73.70	68.07	62.50

Table 4: Results of probing linguistic knowledge through sequential fine-tuning. Accuracy is used as the evaluation metric. A → B means we continue fine-tuning on B after fine-tuning on A.

model by CoLA is forgotten after fine-tuning on SST2.

Moreover, we assume that re-fine-tuning the model on a different task with a larger data size might have a greater impact on the model's linguistic knowledge than fine-tuning on a smaller data size. However, even though the number of samples in SST-2 is much more than CoLA, both SST2 → CoLA and CoLA → SST2 seem to have similar impacts. This indicates that CoLA has similar effects on the linguistic knowledge of BERT fine-tuned on SST-2 as SST-2 has on BERT fine-tuned on CoLA. Therefore, we can conclude that the size of target task data plays a less significant role in impacting the linguistic knowledge obtained during fine-tuning.

Fine-tuning is known to cause models to forget the previously encoded knowledge (Chen et al., 2020), and we witnessed earlier that re-fine-tuning a model does in fact lead a model to forget its knowledge. Hence, we are motivated to ask if and to what extent these "forgotten" knowledge caused by re-fine-tuning is retractable. To answer this, we fine-tune our SST2 → CoLA and CoLA → SST2 models on the task on which they were first fine-tuned. The reported results in Table 4 suggest that regardless of the fine-tuning task and its data size, changes made to a fine-tuned model can be recovered through re-fine-tuning. As an instance, fine-tuning on SST-2 dropped CoordInv performance of the model fine-tuned on CoLA from 69.40 to 65.91. Nevertheless, after re-fine-tuning on CoLA, the accuracy increased to 68.07, which is almost similar to its first state.

Overall, as observed in Table 4, none of the fine-tuning tasks made deep unrecoverable adjustments to the model's linguistic knowledge. We conclude that the changes introduced to the model by fine-tuning are shallow irrespective of the fine-tuning data size.

8 Conclusion

In this paper, we carried out a set of experiments to determine the effects of training data size on the linguistic knowledge captured by fine-tuning. To begin with, by individually probing all layers, we found out that models fine-tuned on larger datasets deviate more from the base model in terms of the encoded linguistic knowledge. We realized that the gaps are more significant in the higher layers, while lower layers possess a similar amount of linguistic knowledge under a fixed data size. As a result, we propose that the comparison of the linguistic knowledge of fine-tuned models is valid if trained on equal data size. Furthermore, the difference in linguistic knowledge across different data sizes can be explained with the number of iterations updating the model during fine-tuning. This suggests that linguistic knowledge is rather affected by the number of fine-tuning iterations than the diversity of the training data. Finally, we discovered that after sequentially fine-tuning on two different downstream tasks, some knowledge is forgotten but can be recovered through re-fine-tuning on the initial downstream task.

We argue that probing accuracy cannot fully represent the linguistic knowledge captured by fine-tuned models, given the fact that a factor, such as size of the dataset, can highly affect probing accuracy. As a future direction, it is crucial to take the undesirable factors that affect probing performance into account. Moreover, evaluating the reliability of existing accuracy-based probes and designing more robust metrics for encoded knowledge assessment are the other important aspects of interpreting LMs.

References

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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)*, pages 4171–4186, 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)*.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957, Online. Association for Computational Linguistics.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Investigating learning dynamics of BERT fine-tuning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 87–92, Suzhou, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. 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.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the Interplay Between Fine-tuning and Sentence-level Probing for Linguistic Knowledge in Pre-trained Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2502–2516, Online. Association for Computational Linguistics.

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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan

Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

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.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

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.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.