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

Houman Mehrafarin[*] [♢] **Sara Rajaee**[*] [♢] **Mohammad Taher Pilehvar** [♠]
[♢]Iran University of Science and Technology, Tehran, Iran
[♠]Tehran Institute for Advanced Studies, Khatam University, Iran
{h_mehrafarin, sara_rajaee}@comp.iust.ac.ir
mp792@cam.ac.uk

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. 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 updating the model during fine-tuning rather than the diversity of the training samples. Finally, we show through a set of experiments that fine-tuning data size affects the recoverability of the changes made to the model's linguistic knowledge.[1]

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 ability 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 higher layers task-specific (Durrani et al., 2021). However, Mosbach et al. (2020) argued that the changes in the probing performance can not be attributed entirely to the modifications a model undergoes with respect to its linguistic knowledge after fine-tuning.

While the previous studies focused on the role of the target task as a factor that affects the probing performance of fine-tuned models, we present another important factor in interpreting probing results for such 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 entirely 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 encoded 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 the extent to which 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

[*]The authors contributed equally to this work.
[1]We have released our code and models' checkpoints at: https://github.com/hmehrafarin/data-size-analysis

Findings of the Association for Computational Linguistics: ACL 2022, pages 228 - 238

layers, particularly for models trained on larger datasets. We take our analysis a step further and show that the difference in probing performance among different data sizes is due to the total number of optimization steps rather than the diversity of training samples. Finally, through a set of experiments, we show that the changes made to the probing performance by a fine-tuning task can be recovered if the model is re-fine-tuned on a task with comparable data size.

The findings of this paper can be summarized as follows:

- Data size is a factor that highly impacts a fine-tuned model's probing performance.
- The size of the dataset mainly affects the probing performance of the higher layers.
- The number of training steps is what makes larger datasets have higher impacts on the model's linguistic knowledge (rather than the diversity in training samples).
- Fine-tuning data size affects the extent to which the modifications made to a model's linguistic knowledge are recoverable.

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 amounts of linguistic features, one still requires to fine-tune them to take full advantage of their 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 the higher layers and alters 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.

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 various 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

	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.

set. We used the [CLS] token for classification and set the learning rate as $5e^{-5}$. 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 presented in Table 1. We report the results on different training data sizes[2] 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.

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

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 allows us to scrutinize the linguistic knowledge by eliminating the possibility of the classifier learning such knowledge. All probes are trained with a batch size of 32, a learning rate of $3e^{-4}$, a linear scheduler for adjusting the learning rate with 10% warm-up steps, and for ten epochs. We also used Adam as the optimizer. Due to limited computational resources, we were not able to run all the experiments multiple times with different random seeds. However, to ensure the reliability of our results, we repeated several randomly chosen experiments three times (with different random seeds). 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[3]. The binary classification tasks are as follows:

Bigram Shift is a task that aims to test the model's ability to predict whether two successive random 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.

[3]We also repeat our experiments on the structural probe of Hewitt and Manning (2019). This probe investigates how well syntactic dependency trees are encoded within a model's representations. We report the UUAS score for the distance between word pairs in the parse tree. The results are reported in Appendix A. We choose this probe because it is different from SentEval's probes in terms of training objective to show our statement still stands.

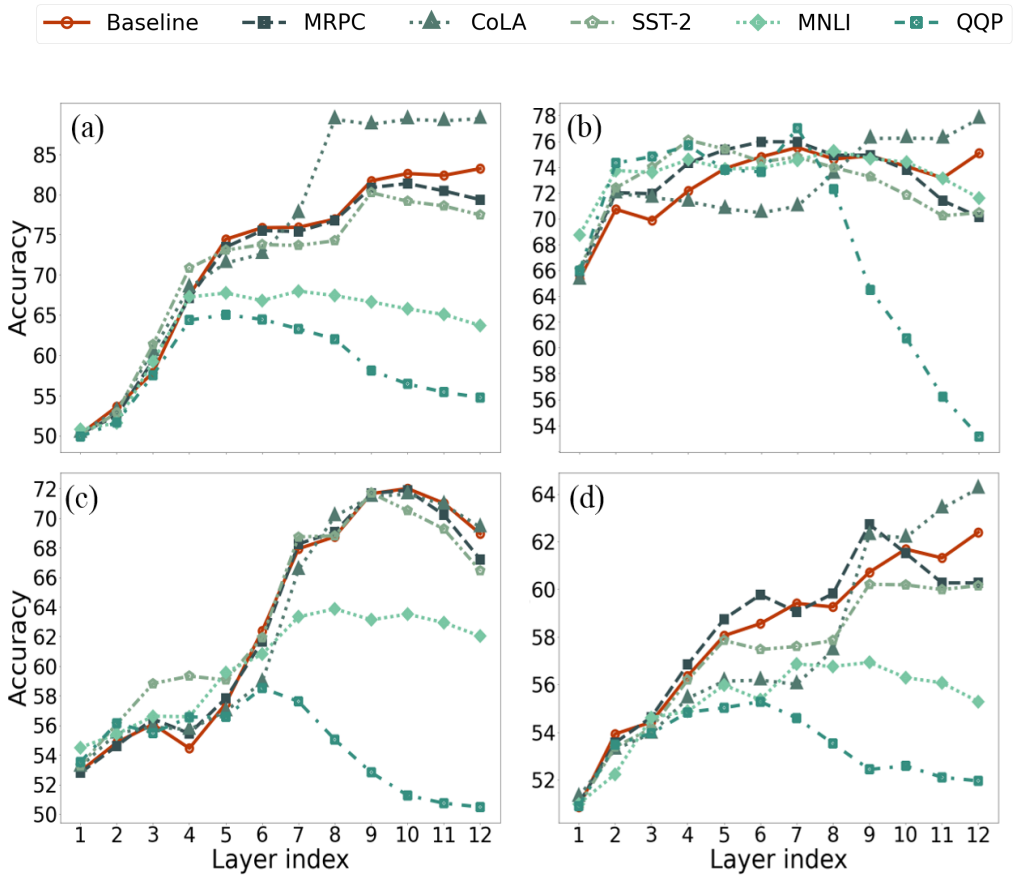


Figure 1: Probing accuracy on all the layers of fine-tuned models on (a) Bigram Shift (b) Object Number (c) Coordination Inversion (d) Semantic Odd Man Out. As shown, there is a large accuracy gap between models fine-tuned on larger data sizes (e.g., MNLI and QQP) and the baseline.

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' size on linguistic knowledge.

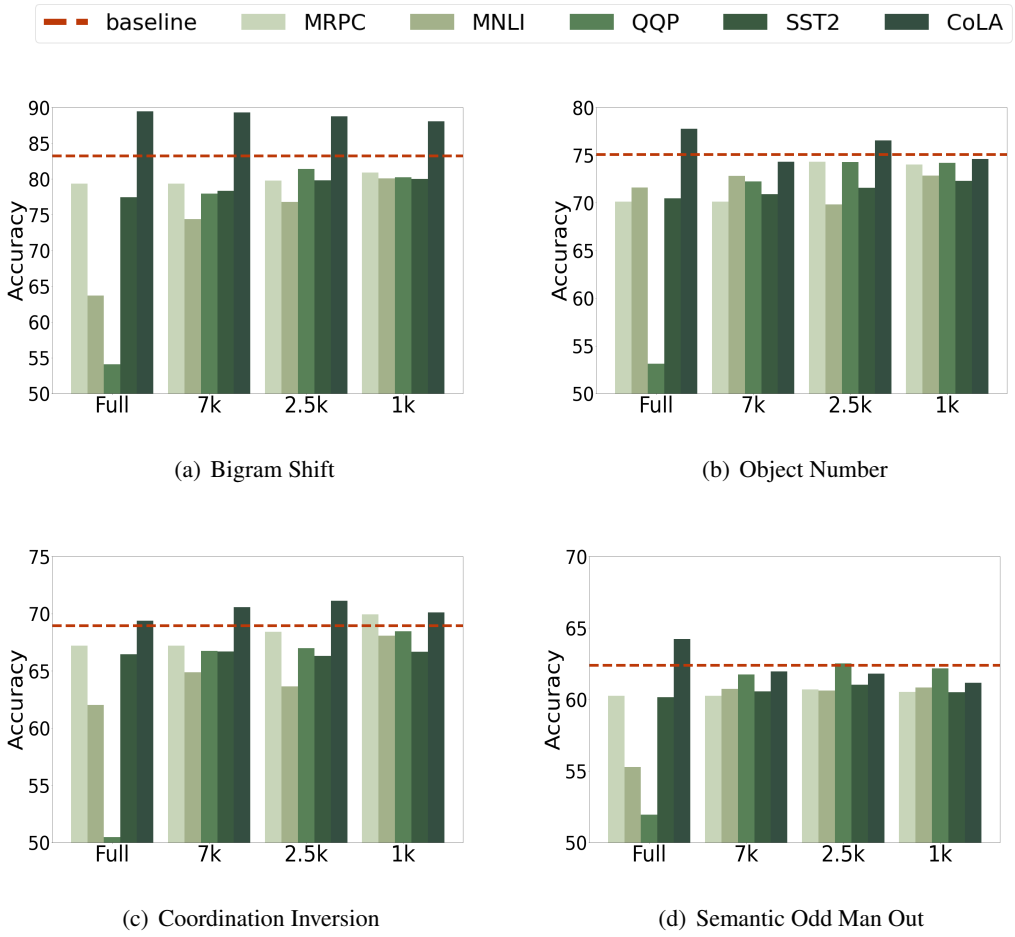
4.1 Probing Linguistic Knowledge

We empirically evaluate the linguistic knowledge captured by several fine-tuned models through the lens of probing performance. 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 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 a smaller 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 highlight the impact of data size on probing accuracy. We can observe that the probing performance of the baseline and models fine-tuned on smaller datasets (e.g., MRPC, SST-2, and CoLA) are comparable, whereas fine-tuning on larger data sizes (e.g., QQP and MNLI) seems to have impacted probing performance by a significant margin. In what follows, we carry out experiments to better understand the reasons behind this observation.

4.2 The Impact of Data Size

One of the popular studies in probing is investigating the changes made to a model's linguistic knowledge after fine-tuning. The changes brought about in the model upon fine-tuning are taken as a means to explain the nature of the corresponding task on which fine-tuning has been carried out (Durrani et al., 2021). Existing studies usually consider several tasks, many of which do not have datasets of



(a) Bigram Shift

(b) Object Number

(c) Coordination Inversion

(d) Semantic Odd Man Out

Figure 2: An illustration of the probing performance of models fine-tuned on fixed-size training sets of five different tasks. The pre-trained BERT's performance on each of the four probing tasks has been shown by the dashed red line. The figures suggest that different fine-tuned models, irrespective of the fine-tuning task, almost encode similar linguistic knowledge when trained on equal-sized data.

comparable size. For instance, in the GLUE benchmark, MNLI is 46 times larger than CoLA. These studies usually focus on the type of downstream tasks only, overlooking the size of their datasets.

Based on our observations in Section 4.1, we hypothesize that, in addition to the type of the downstream task, the size of its corresponding dataset can play an important role in improving or impairing the linguistic knowledge encoded in the model. We examined our hypothesis by fine-tuning pre-trained BERT on the selected downstream tasks with different sets of samples. Specifically, 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 shows the results of this experiment. In general, the results confirm our hypothesis that data size plays a significant role in probing accuracy. In what follows, we further discuss our observations from this experiment.

4.3 Discussion

The effect of data size on both syntactic and semantic probing tasks is notable, denoted by the large gaps between the probing results of the models fine-tuned on larger data sizes and the baseline (see Figure 1). 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), not much change is observed across the results. This shows that fine-tuning data size indeed affects the linguistic knowledge encoded by the model.

Overall, we can conclude that the amount of linguistic knowledge through fine-tuning is highly affected by data size. This suggests that data size

		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
SST-2	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	—	-1.90	-1.41	82.39	-1.05	—	1.36	1.35	61.32
	Layer 12	-3.87	—	-3.45	-2.31	83.23	-2.13	—	-1.70	-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.

should be taken into account when analyzing fine-tuned models.

5 Layer-wise Analysis

Given our observations on the role of data size, we were curious to see how it affects the encoded knowledge in specific layers. As noted by Jawahar et al. (2019), BERT's layers can be divided into three classes in terms of the linguistic knowledge they capture. To this end, we carry out experiments by probing layers 2, 7, and 11-12 to cover all the three categories.

Table 2 shows our results obtained from this experiment, which are compared with BERT-base. Due to our limited resources and the excessive number of experiments, we omitted probing tasks that did not show any distinguishable patterns (Figures 1 and 2), i.e., Coordination Inversion and Object Number. The results follow a similar trend to the ones depicted in Figure 2. As we decrease the number of training samples, the probing performance on the fine-tuned models gets closer to the baseline across all layers. MNLI and QQP's behaviors are compelling evidence of the effectiveness of data size across layers. Such models fine-tuned on larger datasets undergo more considerable changes than those with smaller data sizes.

Regardless of data size, we can also observe that fine-tuning mainly affects higher layers. Our finding is aligned with Merchant et al. (2020) that fine-tuning has a more significant impact on higher layers and negligible effects on lower layers. There is also an interesting pattern concerning CoLA's performance. Despite a drop in performance of around 15% from the full to 1k version (Table 1), the linguistic knowledge has been marginally affected by data size. We leave further investigations on this 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 training data, the gap between the fine-tuned models and the baseline decreases. This behavior can be either attributed to the diversity of training samples or to the larger number of iterations through which the model is updated.

To address this, we repeated the same experiments carried out in Section 5 but with fixing the number of iterations on all data sizes. This allows the model to be fine-tuned for an equal number of iterations across different data sizes of a specific task. Note that we fine-tuned the full models for just one epoch to avoid a large number of iterations for the 7k and 2.5k models. Since SST-2, CoLA, and MRPC have much smaller datasets, and

		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.

the number of iterations does not substantially 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 significantly impairs the captured linguistic knowledge. For instance, we can observe the impact of longer training by comparing Bigram Shift performance on QQP across Tables 2 (54.11) and 3 (73.25)[4]. As Table 3 suggests, fixing the number of iterations reduces the gap across different data sizes, making the 7k and 2.5k models behave almost similarly to the full models. For instance, in Table 2, there is a gap of 24% 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 Linguistic Knowledge Recoverability

Fine-tuning procedure modifies the encoded linguistic knowledge in the pre-trained model. In this section, we aim at verifying the extent to which these modifications are recoverable. To this end, taking a fine-tuned model on a specific task as our baseline, we further fine-tune the model on another task. We then compare the probing performance of the resulting models with their corresponding baselines. High similarity in probing performance indicates the recoverability of the modifications.

We opt for CoLA and SST-2 as a pair of tasks with different linguistic objectives but with comparable data sizes. Also, we experiment with MRPC and QQP, which are similar tasks but with significantly different data sizes (the former's data size is a hundred times larger than the latter's). For instance, considering CoLA and SST-2 as our fine-tuning task pair, *SST-2* $\rightarrow$ *CoLA* $\rightarrow$ *SST-2* stands for a setting where we have consecutively fine-tuned the model on SST-2, CoLA, and SST-2. Following our previous experiments, we report the probing results for the Bigram Shift and Semantic Odd Man Out tasks.[5]

The results are presented in Figure 3. The three-quarters of a circle in the figures represent the maximum value in the corresponding probing task. As shown in the figures, the linguistic knowledge is recoverable through re-fine-tuning on a set of pairs with comparable data sizes. In the previous sections, we observed that CoLA and SST-2 have notably different performances on Bigram Shift and Semantic Odd Man Out. Nevertheless, after re-fine-tuning, both target tasks can recover the knowledge modified by the previous fine-tuning step.

On the other hand, for the QQP and MRPC pair, we observe a different behavior in which the data size of QQP highly limits the extent of knowledge recoverability. Considering Bigram Shift, we observe that the final MRPC fine-tuning in the *QQP* $\rightarrow$ *MRPC* and *MRPC* $\rightarrow$ *QQP* $\rightarrow$ *MRPC* settings can not recover the modification introduced by QQP (the probing results remain similar to QQP's). In the reverse setting (*MRPC* $\rightarrow$ *QQP* and *QQP* $\rightarrow$ *MRPC* $\rightarrow$ *QQP*), the probing performance is negligently affected by MRPC data size, leading to

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

[5] More results for the Object Number and Coordination Inversion tasks can be found in Appendix B.

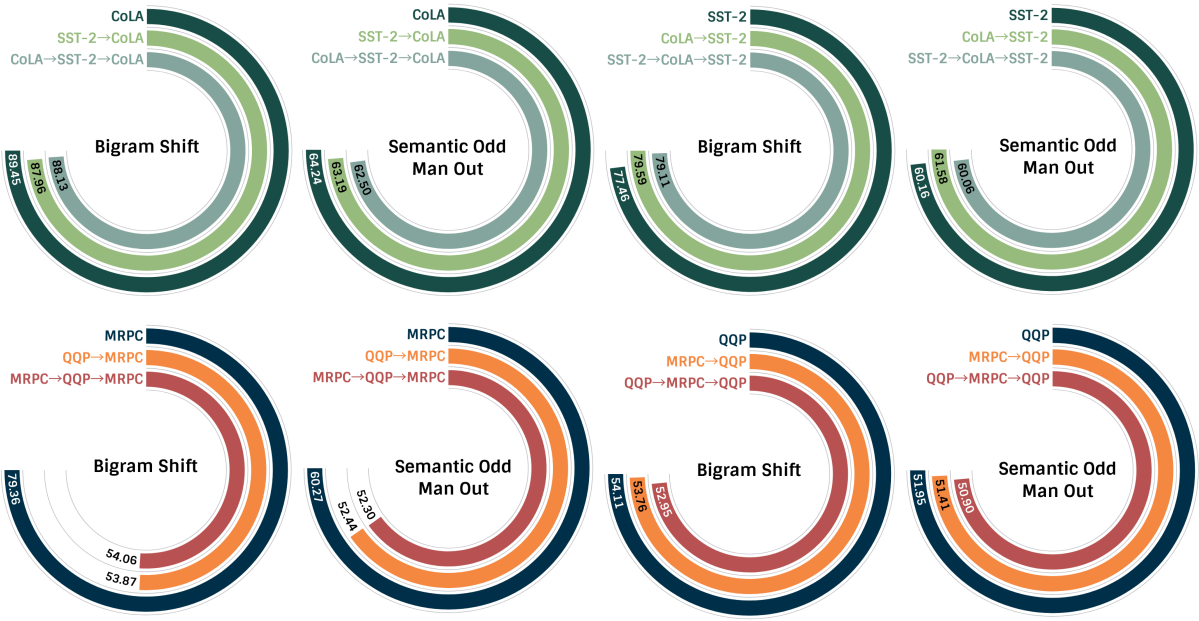


Figure 3: The performance of the models after being sequentially fine-tuned on different tasks. Three-quarters of a circle represents the maximum value and the outer circle is the baseline. The figures demonstrate that the modified knowledge recoverability depends on the fine-tuning data size.

a performance fairly similar to QQP's.[6]

Our results suggest that the extent of knowledge recoverability is bound to the fine-tuning data size. More specifically, further fine-tuning a fine-tuned model with a comparable data size (e.g., *SST-2 → CoLA* and *CoLA → SST-2 → CoLA* introduces the same modifications as fine-tuning a pre-trained model (e.g., CoLA). However, increasing the data size in one of these tasks decreases the extent of recoverability by the other task.

8 Conclusion

In this paper, we carried out a set of experiments to determine the effect of training data size on the probing performance of fine-tuned models. 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 their encoded linguistic knowledge. Therefore, we argue that comparing the linguistic knowledge of fine-tuned models is valid only if they are trained on datasets of comparable sizes. Through layer-wise probing analysis, we realized that the number of training samples mainly affects the probing results for the higher layers, while the results remain similar in the lower layers across different target tasks. Furthermore, we investigated why data size affects the probing performance of fine-tuned models through training the models with limited training data for the same number of iterations as we trained the full models. We showed that the gap in probing performance between models fine-tuned on different data sizes is due to the number of iterations for which the model is updated during fine-tuning rather than the diversity of the training set. Finally, in our last experiment, we showed that the size of a target task's dataset affects the extent to which it can recover the linguistic knowledge previously changed by a different task.

We argue that probing accuracy cannot fully represent the linguistic knowledge captured by fine-tuned models, given that factors, such as the size of the dataset, can highly affect probing accuracy and should be ruled out in any such study. As future work, we plan to evaluate the reliability of existing accuracy and loss-based probes and design more robust metrics for investigating the encoded knowledge in the existing language models.

[6] We have also carried out the exact experiments with QQP 7k to make sure the results are related to the size of the tasks.

References

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.

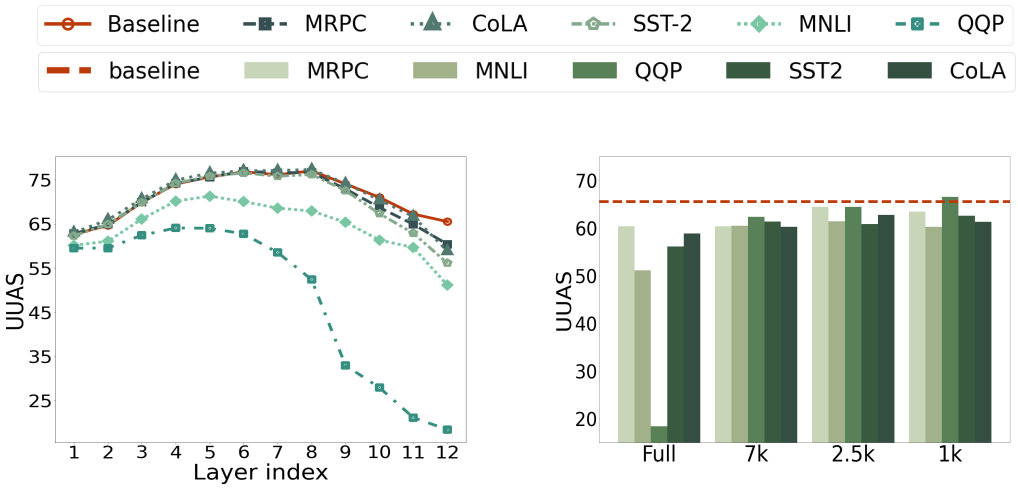
Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

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.

A Structural Probe Analysis

We have also repeated our data size analysis experiment on the structural probe to show that our findings stand for different probes. Figure 4 confirms our conclusions drawn from Section 4, which denotes that data size affects the probing performance of fine-tuned models.

B Linguistic Knowledge Recoverability



(a) Layer-wise analysis (b) Data size impact

Figure 4: (a) UUAS score of the structural probe on all layers of fine-tuned models. (b) The visualization of models' performance fine-tuned on the fixed-size training sets on the structural probe. The pre-trained BERT's performance is shown by the dashed red line.

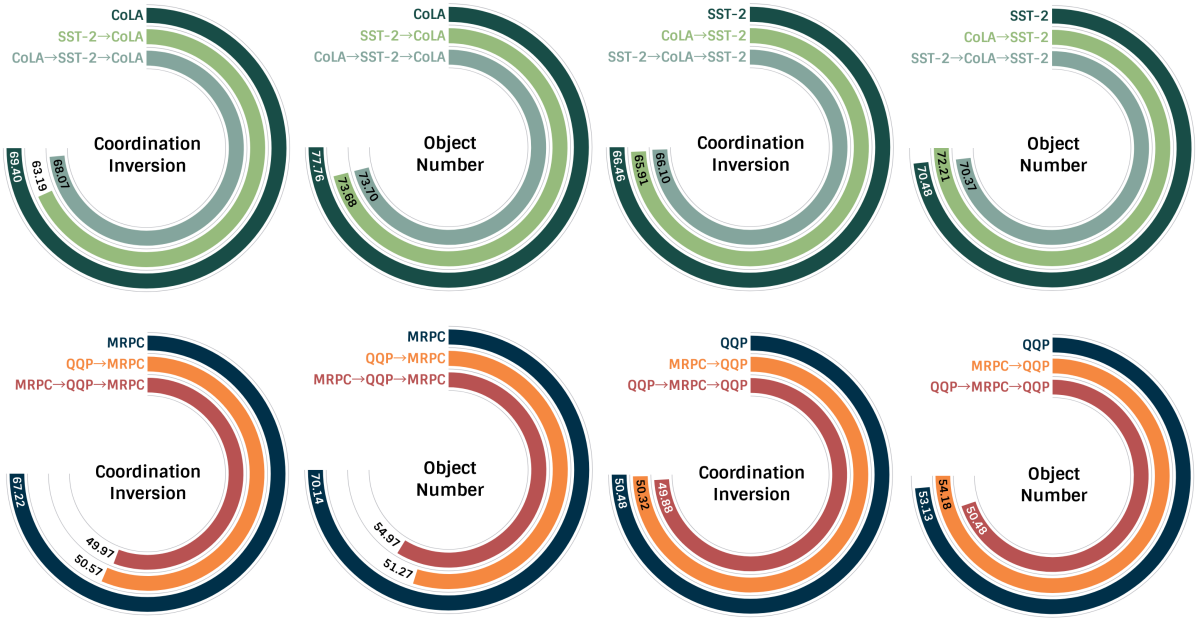


Figure 5: The performance of the models after being sequentially fine-tuned on different tasks. Three-quarters of a circle represents the maximum value and the outer circle is the baseline.