What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning

Abstract

Pretrained transformer-based language models have achieved state of the art across countless tasks in natural language processing. These models are highly expressive, comprising at least a hundred million parameters and a dozen layers. Recent evidence suggests that only a few of the final layers need to be fine-tuned for high quality on downstream tasks. Naturally, a subsequent research question is, "how many of the last layers do we need to fine-tune?" In this paper, we precisely answer this question. We examine two recent pretrained language models, BERT and RoBERTa, across standard tasks in textual entailment, semantic similarity, sentiment analysis, and linguistic acceptability. We vary the number of final layers that are fine-tuned, then study the resulting change in task-specific effectiveness. We show that only a fourth of the final layers need to be fine-tuned to achieve 90% of the original quality. Surprisingly, we also find that fine-tuning all layers does not always help.

Introduction

Transformer-based pretrained language models are a battle-tested solution to a plethora of natural language processing tasks. In this paradigm, a transformer-based language model is first trained on copious amounts of text, then fine-tuned on task-specific data. BERT BIBREF0, XLNet BIBREF1, and RoBERTa BIBREF2 are some of the most well-known ones, representing the current state of the art in natural language inference, question answering, and sentiment classification, to list a few. These models are extremely expressive, consisting of at least a hundred million parameters, a hundred attention heads, and a dozen layers.

An emerging line of work questions the need for such a parameter-loaded model, especially on a single

downstream task. BIBREF3, for example, note that only a few attention heads need to be retained in each layer for acceptable effectiveness. BIBREF4 find that, on many tasks, just the last few layers change the most after the fine-tuning process. We take these observations as evidence that only the last few layers necessarily need to be fine-tuned.

The central objective of our paper is, then, to determine how many of the last layers actually need fine-tuning. Why is this an important subject of study? Pragmatically, a reasonable cutoff point saves computational memory across fine-tuning multiple tasks, which bolsters the effectiveness of existing parameter-saving methods BIBREF5. Pedagogically, understanding the relationship between the number of fine-tuned layers and the resulting model quality may guide future works in modeling.

Our research contribution is a comprehensive evaluation, across multiple pretrained transformers and datasets, of the number of final layers needed for fine-tuning. We show that, on most tasks, we need to fine-tune only one fourth of the final layers to achieve within 10% parity with the full model. Surprisingly, on SST-2, a sentiment classification dataset, we find that not fine-tuning all of the layers leads to improved quality.

Background and Related Work ::: Pretrained Language Models

In the pretrained language modeling paradigm, a language model (LM) is trained on vast amounts of text, then fine-tuned on a specific downstream task. BIBREF6 are one of the first to successfully apply this idea, outperforming state of the art in question answering, textual entailment, and sentiment classification. Their model, dubbed ELMo, comprises a two-layer BiLSTM pretrained on the Billion Word Corpus BIBREF7.

Furthering this approach with more data and improved modeling, BIBREF0 pretrain deep 12- and

24-layer bidirectional transformers BIBREF8 on the entirety of Wikipedia and BooksCorpus BIBREF9. Their approach, called BERT, achieves state of the art across all tasks in the General Language Understanding Evaluation (GLUE) benchmark BIBREF10, as well as the Stanford Question Answering Dataset (BIBREF11).

As a result of this development, a flurry of recent papers has followed this more-data-plus-better-models principle. Two prominent examples include XLNet BIBREF1 and RoBERTa BIBREF2, both of which contest the present state of the art. XLNet proposes to pretrain two-stream attention-augmented transformers on an autoregressive LM objective, instead of the original cloze and next sentence prediction (NSP) tasks from BERT. RoBERTa primarily argues for pretraining longer, using more data, and removing the NSP task for BERT.

Background and Related Work ::: Layerwise Interpretability

The prevailing evidence in the neural network literature suggests that earlier layers extract universal features, while later ones perform task-specific modeling. BIBREF12 visualize the per-layer activations in image classification networks, finding that the first few layers function as corner and edge detectors, and the final layers as class-specific feature extractors. BIBREF13 demonstrate that the low- and high-level notions of content and style are separable in convolutional neural networks, with lower layers capturing content and higher layers style.

Pretrained transformers. In the NLP literature, similar observations have been made for pretrained language models. BIBREF14 analyze BERT's attention and observe that the bottom layers attend broadly, while the top layers capture linguistic syntax. BIBREF4 find that the last few layers of BERT change the most after task-specific fine-tuning. Similar to our work, BIBREF5 fine-tune the top layers of BERT, as part of their baseline comparison for their model compression approach. However, none of the

studies comprehensively examine the number of necessary final layers across multiple pretrained transformers and datasets.

Experimental Setup

We conduct our experiments on NVIDIA Tesla V100 GPUs with CUDA v10.1. We run the models from the Transformers library (v2.1.1; BIBREF15) using PyTorch v1.2.0.

Experimental Setup ::: Models and Datasets

We choose BERT BIBREF0 and RoBERTa BIBREF2 as the subjects of our study, since they represent state of the art and the same architecture. XLNet BIBREF1 is another alternative; however, they use a slightly different attention structure, and our preliminary experiments encountered difficulties in reproducibility with the Transformers library. Each model has base and large variants that contain 12 and 24 layers, respectively. We denote them by appending the variant name as a subscript to the model name.

Within each variant, the two models display slight variability in parameter count—110 and 125 million in the base variant, and 335 and 355 in the large one. These differences are mostly attributed to RoBERTa using many more embedding parameters—exactly 63% more for both variants. For in-depth, layerwise statistics, see Table TABREF4.

For our datasets, we use the GLUE benchmark, which comprises the tasks in natural language inference, sentiment classification, linguistic acceptability, and semantic similarity. Specifically, for natural language inference (NLI), it provides the Multigenre NLI (MNLI; BIBREF16), Question NLI (QNLI; BIBREF10), Recognizing Textual Entailment (RTE; BIBREF17), and Winograd NLI BIBREF18 datasets. For semantic

textual similarity and paraphrasing, it contains the Microsoft Research Paraphrase Corpus (MRPC; BIBREF19), the Semantic Textual Similarity Benchmark (STS-B; BIBREF20), and Quora Question Pairs (QQP; BIBREF21). Finally, its single-sentence tasks consist of the binary-polarity Stanford Sentiment Treebank (SST-2; BIBREF22) and the Corpus of Linguistic Acceptability (CoLA; BIBREF23).

Experimental Setup ::: Fine-Tuning Procedure

Our fine-tuning procedure closely resembles those of BERT and RoBERTa. We choose the Adam optimizer BIBREF24 with a batch size of 16 and fine-tune BERT for 3 epochs and RoBERTa for 10, following the original papers. For hyperparameter tuning, the best learning rate is different for each task, and all of the original authors choose one between 1×10^{-5} and 5×10^{-5}; thus, we perform line search over the interval with a step size of 1×10^{-5}. We report the best results in Table TABREF5.

On each model, we freeze the embeddings and the weights of the first N layers, then fine-tune the rest using the best hyperparameters of the full model. Specifically, if L is the number of layers, we explore $N = \frac{L}{2}, \frac{L}{2} + 1, \dots , L$. Due to computational limitations, we set half as the cutoff point. Additionally, we restrict our comprehensive all-datasets exploration to the base variant of BERT, since the large model variants and RoBERTa are much more computationally intensive. On the smaller CoLA, SST-2, MRPC, and STS-B datasets, we comprehensively evaluate both models. These choices do not substantially affect our analysis.

Analysis ::: Operating Points

We report three relevant operating points in Tables TABREF6–TABREF9: two extreme operating points and an intermediate one. The former is self-explanatory, indicating fine-tuning all or none of the nonoutput

layers. The latter denotes the number of necessary layers for reaching at least 90% of the full model quality, excluding CoLA, which is an outlier.

From the reported results in Tables TABREF6–TABREF9, fine-tuning the last output layer and task-specific layers is insufficient for all tasks—see the rows corresponding to 0, 12, and 24 frozen layers. However, we find that the first half of the model is unnecessary; the base models, for example, need fine-tuning of only 3–5 layers out of the 12 to reach 90% of the original quality—see Table TABREF7, middle subrow of each row group. Similarly, fine-tuning only a fourth of the layers is sufficient for the large models (see Table TABREF9); only 6 layers out of 24 for BERT and 7 for RoBERTa.

Analysis ::: Per-Layer Study

In Figure FIGREF10, we examine how the relative quality changes with the number of frozen layers. To compute a relative score, we subtract each frozen model's results from its corresponding full model. The relative score aligns the two baselines at zero, allowing the fair comparison of the transformers. The graphs report the average of five trials to reduce the effects of outliers.

When every component except the output layer and the task-specific layer is frozen, the fine-tuned model achieves only 64% of the original quality, on average. As more layers are fine-tuned, the model effectiveness often improves drastically—see CoLA and STS-B, the first and fourth vertical pairs of subfigures from the left. This demonstrates that gains decompose nonadditively with respect to the number of frozen initial layers. Fine-tuning subsequent layers shows diminishing returns, with every model rapidly approaching the baseline quality at fine-tuning half of the network; hence, we believe that half is a reasonable cutoff point for characterizing the models.

Finally, for the large variants of BERT and RoBERTa on SST-2 (second subfigure from both the top and

the left), we observe a surprisingly consistent increase in quality when freezing 12–16 layers. This finding suggests that these models may be overparameterized for SST-2.

Conclusions and Future Work

In this paper, we present a comprehensive evaluation of the number of final layers that need to be fine-tuned for pretrained transformer-based language models. We find that only a fourth of the layers necessarily need to be fine-tuned to obtain 90% of the original quality. One line of future work is to conduct a similar, more fine-grained analysis on the contributions of the attention heads.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and enabled by computational resources provided by Compute Ontario and Compute Canada.