Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark

Nikita Nangia[1]
nikitanangia@nyu.edu

Samuel R. Bowman[1,2,3]
bowman@nyu.edu

[1]Center for Data Science
New York University

[2]Dept. of Linguistics
New York University

[3]Dept. of Computer Science
New York University

Abstract

The GLUE benchmark is a suite of language understanding tasks which has seen dramatic progress in the past year, with average performance moving from 70.0 at launch to 82.9, state of the art at writing. Here, we measure human performance on the benchmark, in order to learn whether significant headroom remains for further progress. We provide a conservative estimate of human performance on the benchmark through crowdsourcing: Our annotators are non-experts who must learn each task from a brief set of instructions and 20 examples. In spite of limited training, these annotators robustly outperform the state of the art on six of the nine GLUE tasks and achieve an average score of 87.1. Given the fast pace of progress however, the headroom we observe is quite limited. To reproduce the data-poor setting that our annotators must learn in, we also train the BERT model (Devlin et al., 2019) in limited-data regimes, and conclude that low-resource sentence classification remains a challenge for modern neural network approaches to text understanding.

1 Introduction

This past year has seen tremendous progress in building general purpose models that can learn good language representations across a range of tasks and domains (McCann et al., 2017; Peters et al., 2018; Devlin et al., 2019; Howard and Ruder, 2018; Liu et al., 2019). Reusable models like these can be readily adapted to different language understanding tasks and genres. The General Language Understanding Evaluation (GLUE; Wang et al., 2019) benchmark is designed to evaluate such models. GLUE is built around nine sentence-level natural language understanding (NLU) tasks and datasets, including instances of natural language inference, sentiment analysis, acceptability judgment, sentence similarity, and common sense reasoning.

The recent BigBird model (Liu et al., 2019) —a fine-tuned variant of the BERT model (Devlin et al., 2019)—is state-of-the-art on GLUE at the time of writing, with the original BERT right at its heels. Both models perform impressively enough on GLUE to prompt some increasingly urgent questions: How much better are humans at these NLP tasks? Do standard benchmarks have enough headroom to meaningfully measure further progress? In the case of one prominent language understanding task with a known human performance number, SQuAD 2.0 (Rajpurkar et al., 2018), models built on BERT come extremely close to human performance.[1] On the recent Situations With Adversarial Generations (SWAG; Zellers et al., 2018) dataset, BERT *outperforms* individual expert human annotators. In this work, we estimate human performance on the GLUE test set to determine which tasks see substantial remaining headroom between human and machine performance.

While human performance or interannotator agreement numbers have been reported on some GLUE tasks, the data collection methods used to establish those baselines vary substantially. To maintain consistency in our reported baseline numbers, and to ensure that our results are at least roughly comparable to numbers for submitted machine learning models, we collect annotations using a uniform method for all nine tasks.

We hire crowdworker annotators: For each of the nine tasks, we give the workers a brief training exercise on the task, ask them to annotate a random subset of the test data, and then collect *majority vote* labels from five annotators for each example in the subset. Comparing these labels with the

[1]https://rajpurkar.github.io/SQuAD-explorer/

	Avg	Single Sentence CoLA	SST-2	Sentence Similarity MRPC	STS-B	QQP	Natural Language Inference MNLI	QNLI	RTE	WNLI
Training Size	-	*8.5k*	*67k*	*3.7k*	*7k*	*364k*	*393k*	*108k*	*2.5k*	*634*
Human	**87.1**	**66.4**	**97.8**	80.8/86.3	**92.7/92.6**	80.4/59.5	**92.0/92.8**	91.2	**93.6**	**95.9**
BERT	80.5	60.5	94.9	85.4/89.3	87.6/86.5	**89.3/72.1**	86.7/85.9	92.7	70.1	65.1
BigBird	82.9	62.5	95.6	**88.2/91.1**	89.5/88.8	**89.6/72.7**	86.7/86.0	**94.9**	81.4	65.1
Δ_{bert}	6.6	5.9	2.9	-4.6/-3.0	5.1/6.1	-8.9/-12.6	5.3/6.9	-1.5	23.5	30.8
Δ_{bird}	4.2	3.9	2.2	-7.4/-4.8	3.2/3.8	-9.2/-13.2	5.3/6.8	-3.7	12.2	30.8
Performance on subset with 5-way annotator agreement										
Human	93.7	83.6	100.0	90.2/93.6	98.9/94.7	89.4/74.1	98.5/99.2	95.1	97.4	97.5
BERT	83.5	69.2	97.5	88.9/92.7	95.8/82.3	92.5/78.0	96.4/90.8	93.6	73.0	59.3
Δ	10.2	14.4	2.5	1.3/0.9	3.1/12.4	-3.1/-3.9	2.1/8.4	1.5	24.4	38.2
BERT fine-tuned on less data										
BERT-5000	75.8	57.6	92.0	85.4/89.3	87.1/85.8	82.2/61.0	76.4/76.9	89.2	69.2	65.1
BERT-1000	70.7	49.0	90.4	78.5/84.3	83.6/82.3	77.8/55.8	66.5/68.3	86.6	65.6	65.1
BERT-500	68.5	37.2	88.1	74.0/80.7	77.3/75.2	75.4/51.2	61.8/63.0	85.7	61.5	65.1

Table 1: GLUE test set results. The *Human* baseline numbers are estimated using no more than 500 test examples. All the BERT scores are for BERT-Large. As in the original GLUE paper, we report the Matthews correlation coefficient for CoLA. For MRPC and Quora, we report accuracy then F1. For STS-B, we report Pearson then Spearman correlation. For MNLI, we report accuracy on the matched then mismatched test sets. For all other tasks we report accuracy. The *Avg* column shows the overall GLUE score: an average across each row, weighting each task equally. The Δ_{bert} and Δ_{bird} rows show the difference between the *Human* performance baseline and BERT and BigBird respectively. The second section shows *Human* and BERT performance on the subset of the test set where there is unanimous, 5-way annotator agreement, the Δ row is the difference between them. *Training Size* gives the number of examples in the full training set for each task. The BERT-5000/1000/500 rows show test set results for BERT fine-tuned on no more than 5k, 1k, and 500 examples respectively. Though MRPC and RTE have fewer than 5k examples, we rerun BERT fine-tuning and report these results in the BERT-5000 row.

ground-truth test labels yields an overall GLUE score of 87.1—well above BERT's 80.5 and BigBird's 82.9—and yields single-task scores that are substantially better than both on six of nine tasks. However, in light of the pace of recent progress made on GLUE, the gap in most tasks is relatively small. The one striking exception is the data-poor Winograd Schema NLI Corpus (WNLI; based on Levesque et al., 2012), in which humans outperform machines by over 30 percentage points.

To reproduce the data-poor training regime of our annotators, and of WNLI, we investigate BERT's performance on data-poor versions of the other GLUE tasks and find that it suffers considerably in these low-resource settings. Ultimately however, BERT's performance seems genuinely close to human performance and leaves limited headroom in GLUE.

2 Background and Related Work

GLUE GLUE (Wang et al., 2019) is composed of nine sentence or sentence-pair classification or regression tasks: MultiNLI (Williams et al., 2018), RTE (competition releases 1–3 and 5, merged and treated as a single binary classification task; Dagan et al. 2006, Bar Haim et al. 2006, Giampiccolo et al. 2007, Bentivogli et al. 2009), QNLI (an answer sentence selection task based on SQuAD; Rajpurkar et al. 2016), [2] and WNLI test natural language inference. WNLI is derived from private data created for the Winograd Schema Challenge (Levesque et al., 2012), which specifically tests for common sense reasoning. The Microsoft Research Paraphrase Corpus (MRPC; Dolan and Brockett, 2005), the Semantic Textual Similarity Benchmark (STS-B; Cer et al., 2017), and Quora Question Pairs (QQP)[3] test paraphrase and sentence similarity evaluation. The Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2018) tests grammatical acceptability judgment. Finally, the Stanford Sentiment Treebank (SST;

[2] Our human performance numbers for QNLI are on the original test set since we collected data before the release of the slightly revised second test set. BERT-Large's performance went up by 1.6 percentage points on the new test set, suggesting that our human performance number represents a reasonable—if very conservative—approximation of human performance on QNLI.

[3] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Socher et al., 2013) tests sentiment analysis.

Human Evaluations on GLUE Tasks Warstadt et al. (2018) report human performance numbers on CoLA as well. Using the majority decision from five expert annotators on 200 examples, they get a Matthews correlation coefficient (MCC) of 71.3. Bender (2015) also estimates human performance on the original public Winograd Schema Challenge (WSC) data. They use crowdworkers and report an average accuracy of 92.1%. Wang et al. (2019) report human performance numbers on GLUE's manually curated diagnostic test set. The examples in this test set are natural language inference sentence pairs that are tagged for a set of linguistic phenomena. They use expert annotators and report an average R_3 coefficient of 0.8.

3 Data Collection Method

To establish human performance on GLUE, we hire annotators through the Hybrid[4] data collection platform, which is similar to Amazon's Mechanical Turk. Each worker first completes a short training procedure then moves on to the main annotation task. For the annotation phase, we tune the pay rate for each task, with an average rate of $17/hr. The training phase has a lower, standard pay rate, with an average pay of $7.6/hr.

Training In the training phase for each GLUE task, each worker answers 20 random examples from the task development set. Each training page links to instructions that are tailored to the task, and shows five examples. Their answers for these examples can be revealed by clicking on a "Show" button at the bottom of the page. We ask the workers to label each set of examples and check their work so they can familiarize themselves with the task. Workers who get less than 65% of the examples correct during training do not qualify for the main task. This is an intentionally low threshold meant only to encourage a reasonable effort. Our platform cannot fully prevent workers from changing their answers after viewing the correct labels, so we cannot use the training phase as a substantial filter. (See Appendix A.1 in the supplement for details on the training phase.)

Annotation We randomly sample 500 examples from each task's test set for annotation, with the exception of WNLI where we sample 145 of the 147 available test examples (the two missing examples are the result of a data preparation error). For each of these sampled data points, we collect five annotations from five different workers (see Appendix A.2). We use the test set since the test and development sets are qualitatively different for some tasks, and we wish to compare our results directly with those on the GLUE leaderboard.

[4] http://www.gethybrid.io

4 Results and Discussions

To calculate the human performance baseline, we take the majority vote across the five crowdsourced annotations. In the case of MultiNLI, since there are three possible labels—*entailment*, *neutral*, and *contradiction*—about 2% of examples see a tie between two labels. For these ties, we take the label that is more frequent in the development set. In the case of STS-B, we take an average of the scalar annotator labels. Since we only collect annotations for a subset of the data, we cannot access the test set through the GLUE leaderboard interface, we instead submit our predictions to the GLUE organizers privately.

We compare human performance to BERT and BigBird. The human performance numbers in Table 1 shows that overall our annotators *stick it to the Muppets* on GLUE. However on MRPC, QQP, and QNLI, Bigbird and BERT *outperform* our annotators. The results on QQP are particularly surprising: BERT and BigBird score over 12 F1 points better than our annotators. Our annotators, however, are only given 20 examples and a short set of instructions for training, while BERT and BigBird are fine-tuned on the 364k-example QQP training set. In addition, we find it difficult to compose concise instructions for QQP that actually match the supplied labels. We do not have access to the material used to create the dataset, and we find it difficult to intfer simple instructions from the data (sample provided in Appendix B). If given more training data, it is possible that our annotators could better learn relatively subtle label definitions that better fit the corpus.

Unanimous Vote To investigate the possible effect of ambiguous label definitions, we look at human performance when there is 5-way annotator agreement. Using unanimous agreement, rather than majority agreement, has the potential effect of filtering out examples of two kinds: those for which our supplied annotation guidelines don't provide clear advice and those for which humans

understand the expectations of the task but find the example genuinely difficult or uncertain. To disentangle the two effects, we also look at BERT results on this subset of the test set, as BERT's use of large training sets means that it should only suffer in the latter cases. We get consent from the authors of BERT to work in cooperation with the GLUE team to measure BERT's performance on this subset, which we show in Table 1. Overall, we see the gap widen between the human baseline and BERT by 3.1 points. The largest shifts in performance are on CoLA, MRPC, QQP, and WNLI. The relative jumps in performance on MRPC and QQP support the claim that human performance is hurt by imprecise guidelines and that the use of substantially more training data gives BERT an edge on our annotators.

In general, BERT needs large datasets to fine-tune on. This is further evidenced by its performance discrepancy between MultiNLI and RTE: human performance is similar for the two, whereas BERT shows a 16.2 percentage point gap between the two datasets. Both MultiNLI and RTE are textual entailment datasets, but MultiNLI's training set is quite large at 393k examples, while the GLUE version of RTE has only 2.5k examples. However, BigBird does not show as large a gap, which may be because it employs a multi-task learning approach which fine-tunes the model for all sentence-pair tasks jointly. Their RTE classifier, for example, benefits from the large training dataset for the closely related MultiNLI task.

Low-Resource BERT Baseline To understand the impact of abundant target task on the limited headroom that we observe, we train several additional baselines. In these, fine-tune BERT on 5k, 1k, and 500 examples for each GLUE task (or fewer for tasks with fewer training examples). We use BERT for this analysis because the authors have released their code and have provided pretrained weights for the model. We use their publicly available implementation of BERT-Large, their pretrained weights as the initialization for fine-tuning on the GLUE tasks, and the hyperparameters they report. We see a precipitous drop in performance on most tasks with large datasets, with the exception of QNLI. A possible partial explanation is that both QNLI and the BERT training data come from English Wikipedia. On MRPC and QQP however, BERT's performance drops below human performance in the 1k- and 500-example settings. On the whole, we find that BERT suffers in low-resource settings. These results are in agreement with the findings in Phang et al. (2019) who conduct essentially the same experiment.

CoLA Our human performance number on CoLA is 4.9 points below what was reported in Warstadt et al. (2018). We believe this discrepancy is because they use linguistics PhD students as expert annotators while we use crowdworkers. This further supports our belief that our human performance baseline is a conservative estimate, and that higher performance is possible, particularly with more training.

WNLI No system on the GLUE leaderboard has managed to exceed the performance of the most-frequent-class baseline on WNLI, and several papers that propose methods for GLUE justify their poor performance by asserting that the task must be somehow broken.[5] WNLI's source Winograd Schema data was constructed so as not to include any statistical cues that a simple machine learning system can exploit, which can make it quite difficult. The WNLI test set shows one of the *highest* human performance scores of the nine GLUE tasks, reflecting its status as a corpus constructed and vetted by artificial intelligence experts. This affirms that tasks like WNLI with small training sets (634 sentence pairs) and no simple cues remain a serious (and sometimes unacknowledged) blind spot for modern neural network sentence understanding methods.

5 Conclusion

This paper presents a conservative estimate of human performance to serve as a target for the GLUE sentence understanding benchmark. We obtain this baseline with the help of crowdworker annotators. We find that state-of-the-art models like BERT are not far behind human performance on most GLUE tasks. But we also note that, when trained in low-resource settings, BERT's performance falls considerably. Given these results, and the continued difficulty neural methods have with the Winograd Schema Challenge, we argue that future work on GLUE-style sentence understanding tasks might benefit from a focus on learning from smaller training sets.

[5] Devlin et al. (2019), for example, mention that they avoid "the problematic WNLI set".

References

Roy Bar Haim, Ido Dagan, Bill Dolan, Ferro Lisa, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *MAICS*.

Luisa Bentivogli, Ido Dagan, Hao Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of Text Analysis Conference*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo 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. Association for Computational Linguistics.

Ido Dagan, Oren Glickmen, and Magnini Bernardo. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges*, pages 177–190. Springer.

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*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *In Proceedings of IWP*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *ACL-PASCAL workshop on textual entailmentand paraphrasing*, pages 1–9. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint 1901.11504*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

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*. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint 1811.01088*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. 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. 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*. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *arXiv preprint 1805.12471*.

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*, pages 1112–1122. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.