1 GLUE benchmark information

Recent NLP models are being evaluated and compared according to their scores in several among GLUE benchmarks[1] ,[2] including a collection of multiple different tasks that we highlight in table 1 such as sentiment (SST), paraphrase (MRPC, QQP), sentence similarity (STS), coreference, question answering (QNLI), textual entailment (RTE, MNLI), etc., where evaluation is performed by utilizing different metrics, such as accuracy, F$_1$ score, Matthews, Pearson, Spearman correlation measures etc. for each different benchmark (1).

Corpus	**|Train|**	**|Test|**	**Task**	**Metrics**	**Domain**
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: **Task descriptions and statistics.** All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form. Table extracted from (1).

		Single Sentence		**Similarity and Paraphrase**			**Natural Language Inference**			
Model	**Avg**	**CoLA**	**SST-2**	**MRPC**	**QQP**	**STS-B**	**MNLI**	**QNLI**	**RTE**	**WNLI**
					Recurrent Neural Networks					
BiLSTM	63.9	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	75.7	52.8	**65.1**
+ELMo	66.4	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	71.7	50.1	**65.1**
+CoVe	64.0	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	75.4	53.5	**65.1**
+Attn	63.9	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	77.2	51.9	**65.1**
+Attn, ELMo	66.5	**35.0**	90.2	68.8/80.2	**86.5**/**66.1**	55.5/52.5	**76.9**/**76.7**	76.7	50.4	**65.1**
+Attn, CoVe	63.2	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	74.5	52.7	**65.1**
					Transformers					
BERT$_{\text{BASE}}$	79.6	52.1	93.5	88.9	71.2	85.8	84.6/83.4	90.5	66.4	N/A
BERT$_{\text{LARGE}}$	**82.1**	**60.5**	**94.9**	**89.3**	**72.1**	**86.5**	**86.7**/**85.9**	**92.7**	**70.1**	N/A

Table 2: **Test performance on GLUE benchmarks (Single-Task Training)**. For CoLA, we report Matthews correlation. For all other tasks we report accuracy. Table partly extracted from (1) and (2) for the respective models.

References

[1] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," 2019.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[1] The python script used to download data for GLUE benchmarks is https://github.com/georgmosh/gluescript

[2] General Language Understanding Evaluation (GLUE) benchmark is a collection of resources used for training, evaluating, and analyzing NLP systems, https://gluebenchmark.com/