A The GLUE and SciTail Datasets

Basically, the GLUE dataset (Wang et al., 2019) consists of three types of tasks: single-sentence classification, similarity and paraphrase tasks, and inference tasks, as shown in Table 1.

Single-Sentence Classification. The model needs to make a prediction given a single sentence for this type of tasks. The goal of the **CoLA** task is to predict whether an English sentence is grammatically plausible and the goal of the **SST-2** task is to determine whether the sentiment of a sentence is positive or negative.

Similarity and Paraphrase Tasks. For this type of tasks, the model needs to determine whether or to what extent two given sentences are semantically similar to each other. Both the **MRPC** and the **QQP** tasks are classification tasks that require the model to predict whether the sentences in a pair are semantically equivalent. The **STS-B** task, on the other hand, is a regression task and requires the model to output a real-value score representing the semantic similarity of the two sentences.

Inference Tasks. Both the **RTE** and the **MNLI** tasks aim at predicting whether a sentence is entailment, contradiction or neutral with respect to the other. **QNLI** is converted from a question answering dataset, and the task is to determine whether the context sentence contains the answer to the question. **WNLI** is to predict if the sentence with the pronoun substituted is entailed by the original sentence. Because the test set is imbalanced and the development set is adversarial, so far none of the proposed models could surpass the performance of the simple majority voting strategy. Therefore, we do not use the WNLI dataset in this paper.

SciTail is a textual entailment dataset that is derived from a science question answering dataset (Khot et al., 2018). Given a premise and a hypothesis, the model need to determine whether the premise entails the hypothesis. The dataset is fairly difficult as the sentences are linguistically challenging and the lexical similarity of premise and hypothesis is high.

B Implementation Details

Our implementation is based on the PyTorch implementation of BERT.[1] We first load the pretrained **BERT$_{\textbf{BASE}}$** model. We use the Adam optimizer (Kingma and Ba, 2015) with a batch size of 32 for both meta-learning and fine-tuning. We set the maximum length to 80 to reduce GPU memory usages.

In the meta-learning stage, we use a learning rate of 5e-5 to train the models for 5 epochs. Both the dropout and the warm-up ratio are set to 0.1 and we do not use gradient clipping. We set the update step k to 5, the number of sampled tasks in each step to 8 and α to 1e-3.

For fine-tuning, again the dropout and warum-up ratio are set to 0.1 and we do not use gradient clipping. The learning rate is selected from {5e-6, 1e-5, 2e-5, 5e-5} and the number of epochs is selected from {3, 5, 10, 20}. We select hyperparameters that achieve the best performance on the development set.

We do not use the stochastic answer network as in MT-DNN for efficiency.

C Linguistic Information

In this part, we use 10 probing tasks (Conneau et al., 2018) to study what linguistic information is captured by each layer of the models.

A probing task is a classification problem that requires the model to make predictions related to certain linguistic properties of sentences. The abbreviations for the 10 tasks are listed in Table 2. Basically, these tasks are set to test the model's abilities to capture surface, syntactic or semantic information. We refer the reader to Conneau et al. (2018) for details. We freeze all the parameters of the models and only train the classification layer for the probing tasks.

First, we can see that the BERT model captures more surface, syntactic and semantic information than other models, suggesting it learns more general representations. MT-DNN and our models, on the other hand, learn representations that are more tailored to the GLUE tasks.

Second, our models perform better than MT-DNN on the probing tasks, indicating meta-learning algorithms may find a balance between general linguistic information and task-specific information. Among the three meta-learning algo-

[1] https://github.com/huggingface/pytorch-pretrained-BERT

Corpus	**Task**	**# Train**	**# Label**	**Metrics**
Single-Sentence Tasks				
CoLA	Acceptability	8.5k	2	Matthews correlation
SST-2	Sentiment	67k	2	Accuracy
Similarity and Paraphrase Tasks				
MRPC	Paraphrase	3.7k	2	F1/Accuracy
STS-B	Similarity	7k	1	Pearson/Spearman correlation
QQP	Paraphrase	364k	2	F1/Accuracy
Inference Tasks				
MNLI	NLI	393k	3	Accuracy
QNLI	QA/NLI	105k	2	Accuracy
RTE	NLI	2.5k	2	Accuracy
WNLI	NLI	634	2	Accuracy
SciTail	NLI	23.5k	2	Accuracy

Table 1: Basic information and statistics of the GLUE and SciTail datasets (Williams et al., 2018).

Model	**Surface**		**Syntactic**				**Semantic**			
	SentLen	Word	TreeDep	ToCo	BShif	Tense	SubNum	ObjNum	SOMO	CoIn
Majority Voting	16.67	0.10	17.88	5.00	50.00	50.00	50.00	50.00	50.13	50.00
BERT-Layer 1	90.84	7.54	32.31	57.91	50.67	78.83	77.50	75.65	50.13	50.05
BERT-Layer 6	69.87	1.06	31.96	76.97	79.66	86.19	84.33	77.58	57.73	63.43
BERT-Layer 12	63.15	32.98	28.80	71.36	85.67	89.72	76.63	76.52	60.92	70.91
MTDNN-Layer 1	92.43	25.84	33.57	58.64	50.00	78.00	80.70	79.83	51.26	51.57
MTDNN-Layer 6	80.11	21.41	31.73	59.58	76.00	81.89	80.36	80.00	55.52	58.31
MTDNN-Layer 12	58.15	23.49	28.03	56.93	75.58	85.47	76.94	72.76	58.16	66.09
MAML-Layer 1	92.21	2.09	30.64	55.27	50.00	77.71	72.61	70.44	50.13	52.49
MAML-Layer 6	76.26	32.13	28.24	67.45	68.43	87.88	80.79	80.07	55.40	59.38
MAML-Layer 12	61.50	20.32	27.31	60.15	79.47	85.56	77.60	75.86	56.76	63.59
FOMAML-Layer 1	88.39	3.22	30.91	51.01	49.97	79.56	74.53	71.28	50.13	50.00
FOMAML-Layer 6	81.33	22.63	30.44	69.48	77.01	88.89	81.81	80.18	57.93	60.11
FOMAML-Layer 12	62.93	30.84	28.33	59.15	79.96	87.60	79.33	77.98	58.05	64.58
Reptile-Layer 1	87.97	3.26	30.00	52.88	50.74	80.48	74.32	70.90	50.13	50.00
Reptile-Layer 6	77.55	24.52	30.74	69.18	75.20	88.42	82.11	81.03	58.52	61.39
Reptile-Layer 12	60.02	29.07	27.78	59.00	82.95	87.34	77.75	75.21	59.23	67.60

Table 2: Accuracy numbers on the 10 probing tasks (Conneau et al., 2018).

rithms, Reptile can capture more general linguistic information than others. Considering Reptile has outperformed the other two models on the GLUE dataset, these results further demonstrate Reptile may be more suitable for NLU tasks.

Third, we find that there may not always exist a monotonic trend on what linguistic information each layer captures. Also, contrary to the findings from Liu et al. (2019) which suggest the middle layers of BERT are more transferable and contain more syntactic and semantic information, our experimental results demonstrate that this may not always be true. We conjecture this is because both syntactic and semantic information are broad concepts and the probing tasks in Liu et al. (2019) may not cover all of them. For example, there exist a monotonic trend for SOMO while the middle layers of these models are better at tasks like SubNum.

Another interesting thing to note is that the lower layers of models perform rather poorly on the word content task, which tests whether the model can recover information about the original words in the sentence. We attribute this phenomenon to the use of subwords and position/token embeddings. In the higher layers, the model may gain more word-level information through the self-attention mechanism.

References

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single\ &!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *AAAI*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *NAACL*.

Alex Wang, Amapreet 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 *ICLR*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.