BERT and PALs: Supplementary Material

A. Performance on Tasks Over Time

Figure 1 shows performance on the GLUE tasks over time for PALs and low-rank adapter modules. The low-resource tasks have a much larger variation in performance than the high resource ones, which are fairly stable. CoLA performance in particular varies a lot early on in training. Performance on CoLA and RTE goes down towards the end of training with low-rank adapters, and not with PALs, and the opposite trend for MRPC. These downward trends might be rectified with a better training schedule or regularisation scheme.

B. Squad and SWAG Performance

We conducted limited experiments on two additional tasks. The Stanford Question Answering Dataset (SQuAD) is a collection of 100k crowdsourced question/answer pairs (Rajpurkar et al., 2016), where the task is to predict the location of the answer in a paragraph from Wikipedia. We follow the approach of Devlin et al. (2018) by associating each token in the input sequence with a probability of being the start, and end, of the answer span. The Situations With Adversarial Generations (SWAG) dataset contains 113k sentence-pair completion examples intended to evaluate grounded commonsense inference (Zellers et al., 2018). Given a sentence from a video captioning dataset, the task is to decide among four choices the most plausible continuation, with each sentence-completion pair assigned a score, and a softmax applied over the four choices to form a probability distribution.

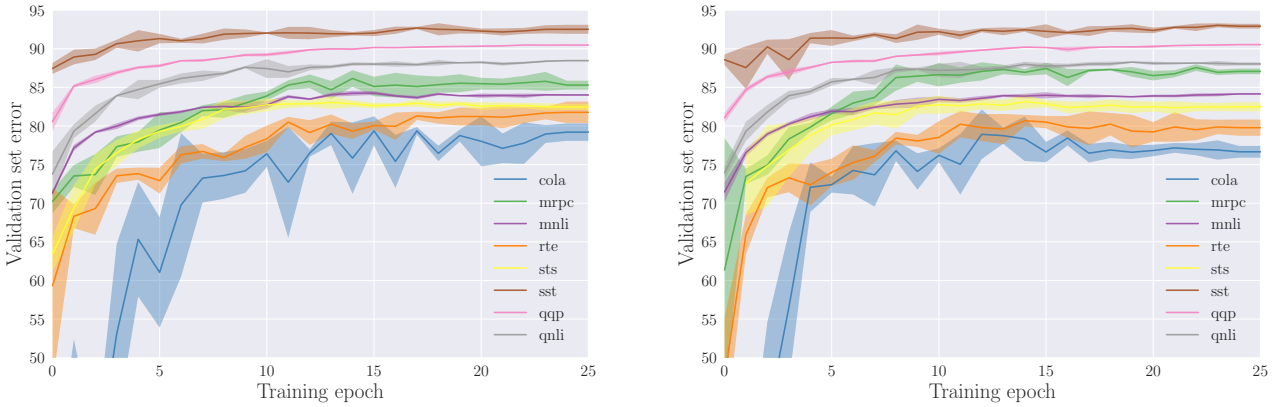
We tested multi-task learning with the SQuAD and SWAG datasets. We follow all the same experimental settings as before, but we use round robin sampling because of the comparable size of the datasets, and train for 24,000 steps, not 60,000, with an increased maximum sequence length, 256. Results, see table 1, show a slight improvement when using the PAL adapters compared to a fully shared baseline and low-rank adapters. However all approaches performed similarly, with there perhaps less need for the flexibility provided by adapters when only training on two tasks.

References

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1264.

Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.



(a) PALs (b) Low rank adapters

Figure 1. Average performance over four random seeds for two adapter modules, with the shaded region indicating standard deviation. CoLA performance has been shifted up by 30% for visibility.

Table 1. Performance on SQuAD and SWAG, in terms of average score across each task's development set; this score is exact match and f1 score for SQuAD, and accuracy for SWAG.

METHOD	NO. PARAMS	NEW LAYERS	ROUND ROBIN
SHARED	1.00×	0	82.75±0.09
ADDING WITHIN BERT			
PALS (204)	1.13×	12	82.774±0.006
LOW RANK (100)	1.13×	12	82.74±0.06