Parameter-Efficient Transfer Learning for NLP

N. Houlsby, A. Giurgiu*, S. Jastrzębski*, B. Morrone,
Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly

Google AI

Imagine doing Transfer Learning for NLP

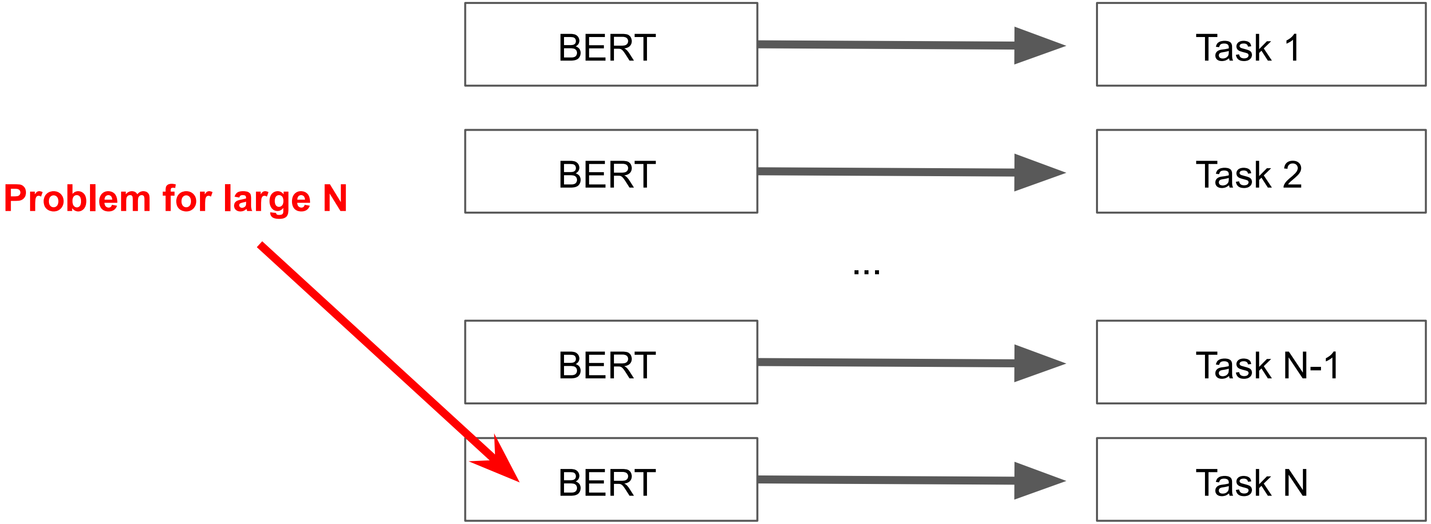
Ingredients:

- A large pretrained model (BERT)
- Fine-tuning

Imagine doing Transfer Learning for NLP

Ingredients:

- A large pretrained model (BERT)
- Fine-tuning



BERT → Task 1

BERT → Task 2

...

BERT → Task N-1

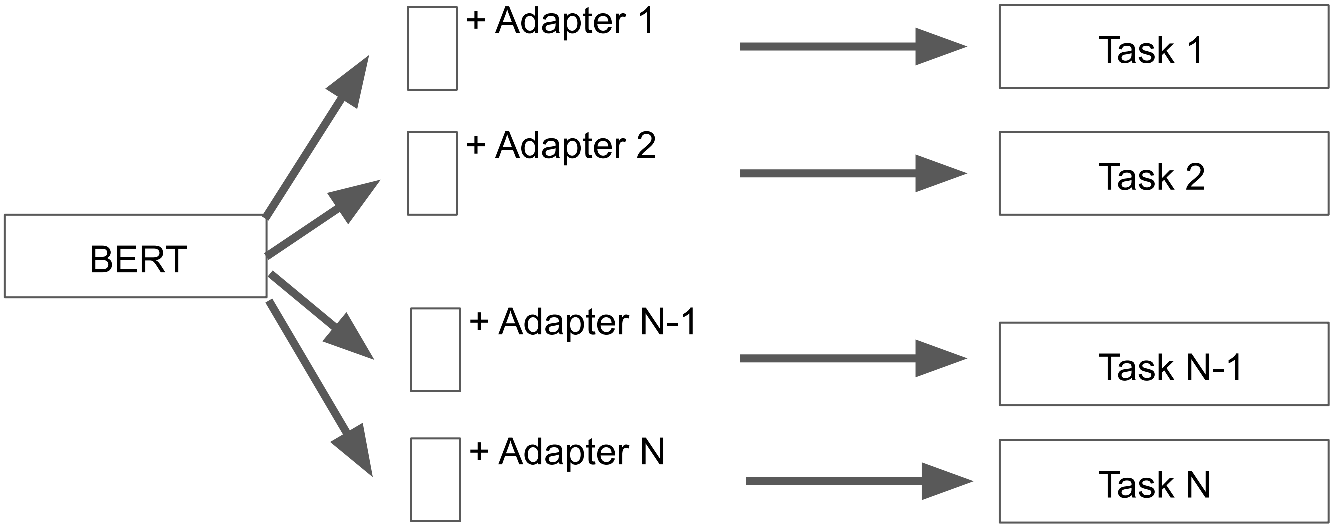
BERT → Task N

Problem for large N

Imagine doing Transfer Learning for NLP

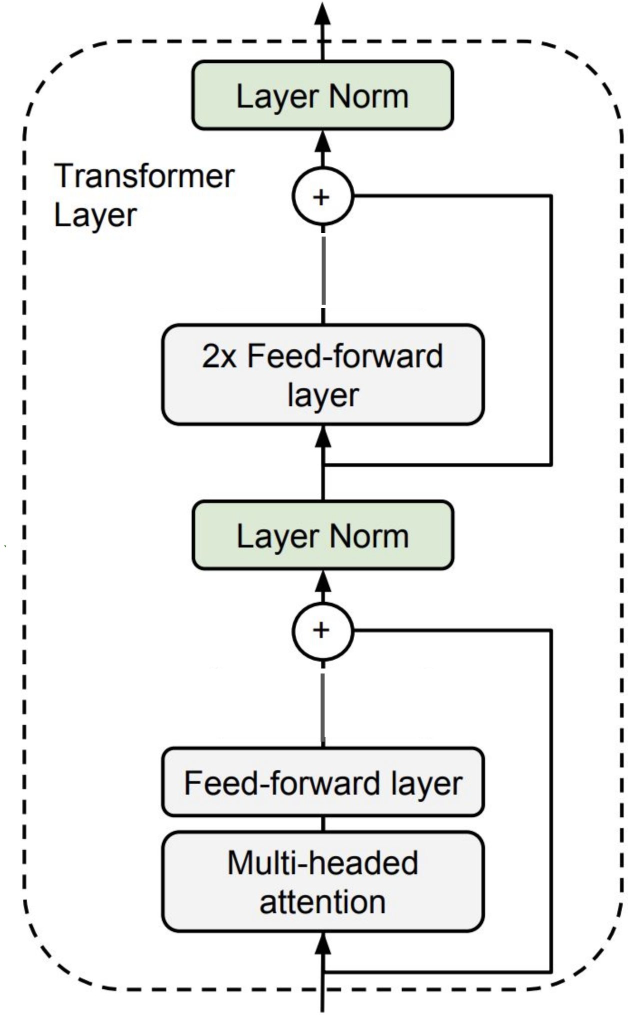
Ingredients:

- A large pretrained model (BERT)
- Fine-tuning



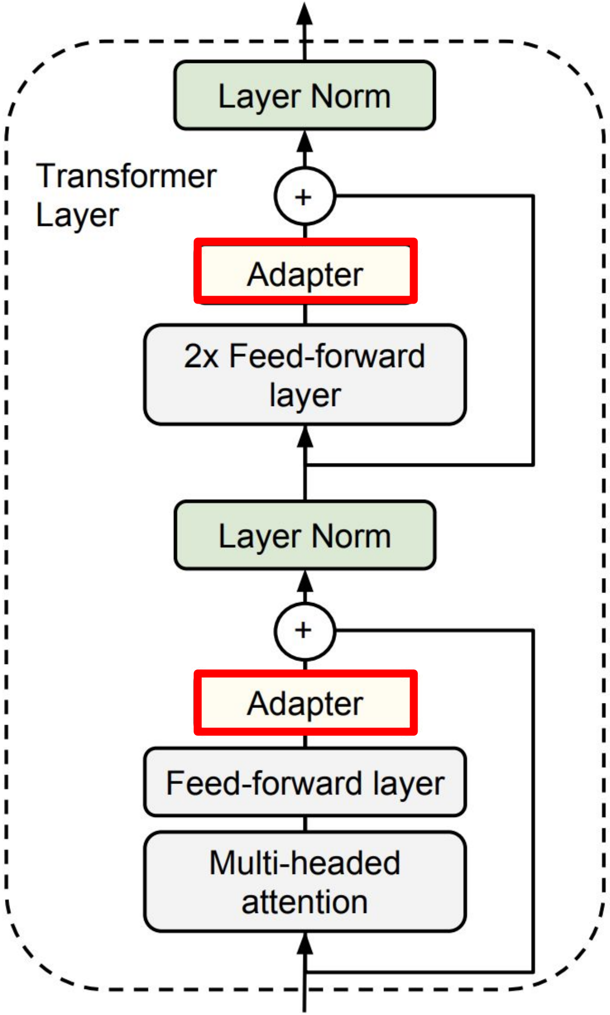
BERT + Adapters

- **Solution**: Train tiny adapter modules at each layer



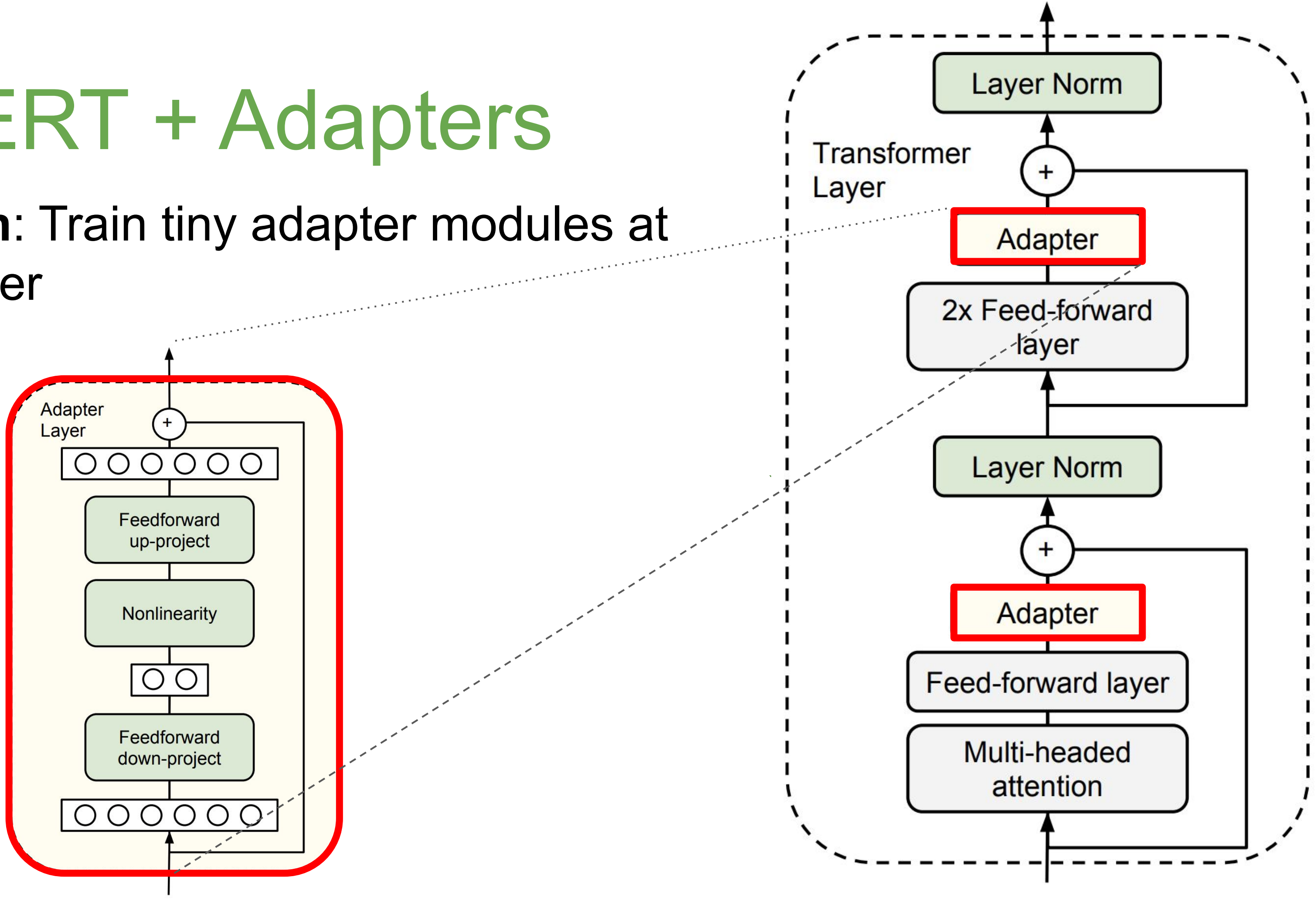
BERT + Adapters

- **Solution**: Train tiny adapter modules at each layer



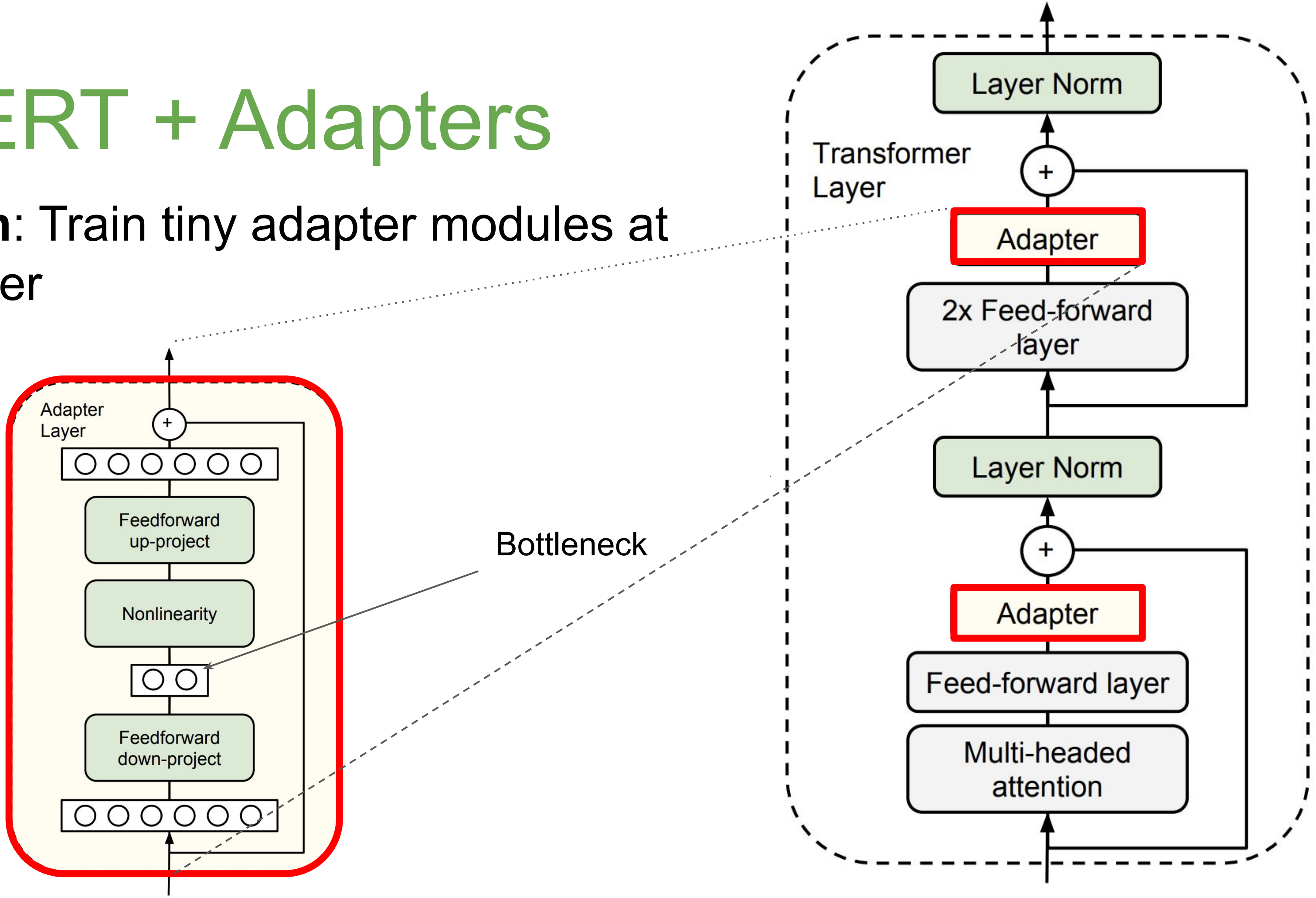
BERT + Adapters

- **Solution**: Train tiny adapter modules at each layer

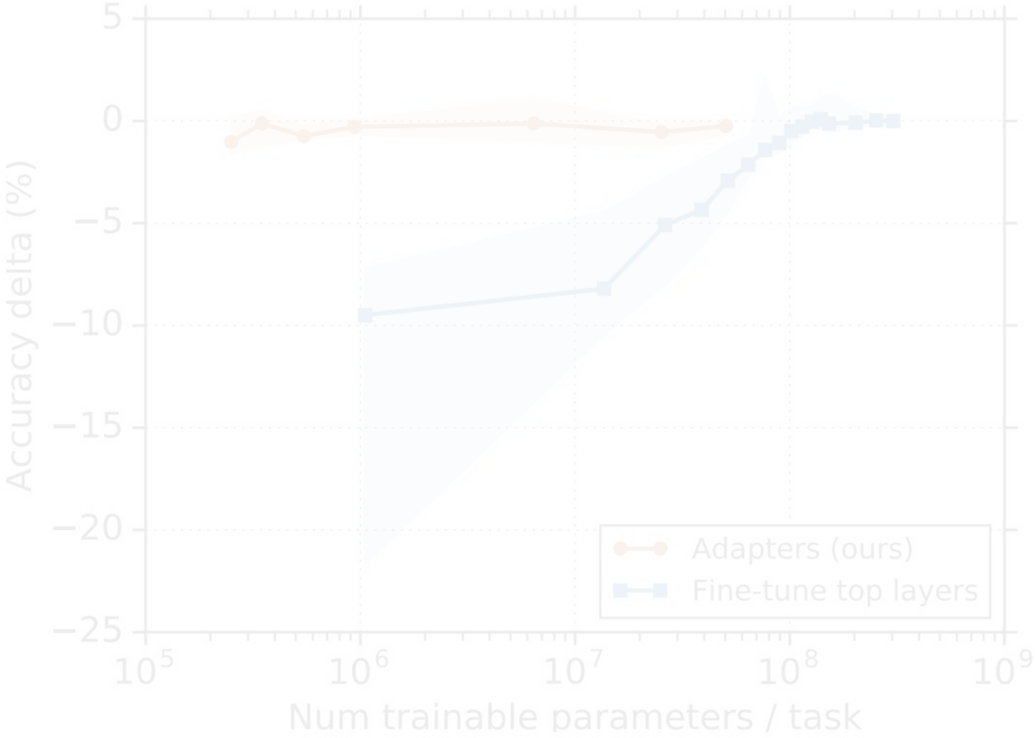


BERT + Adapters

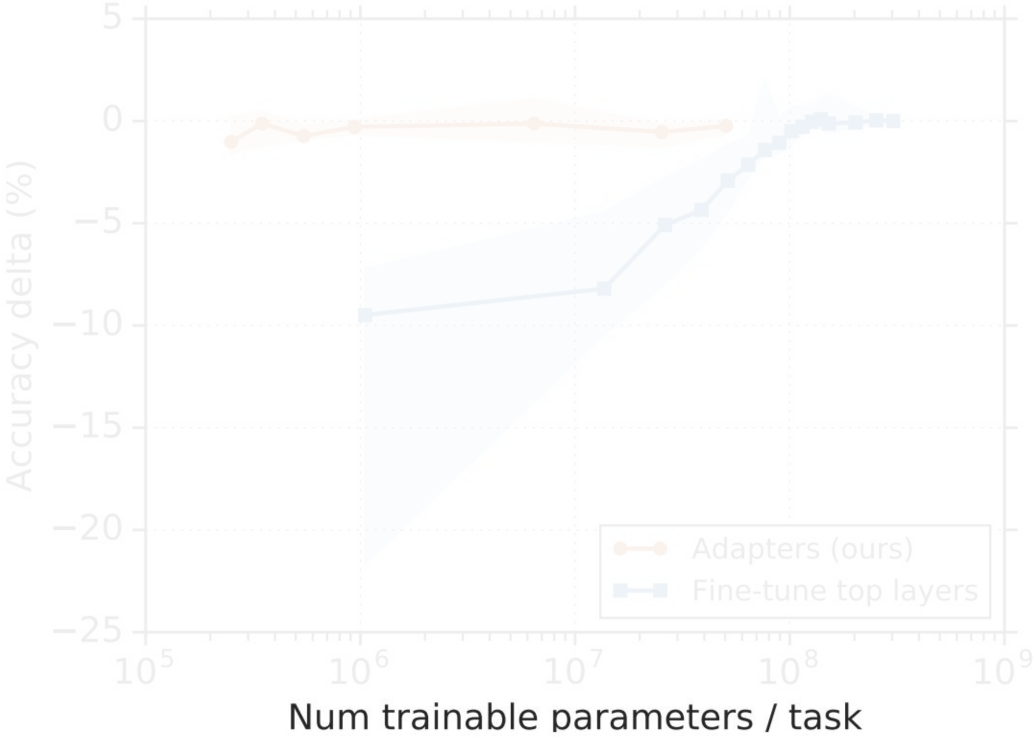
- **Solution**: Train tiny adapter modules at each layer



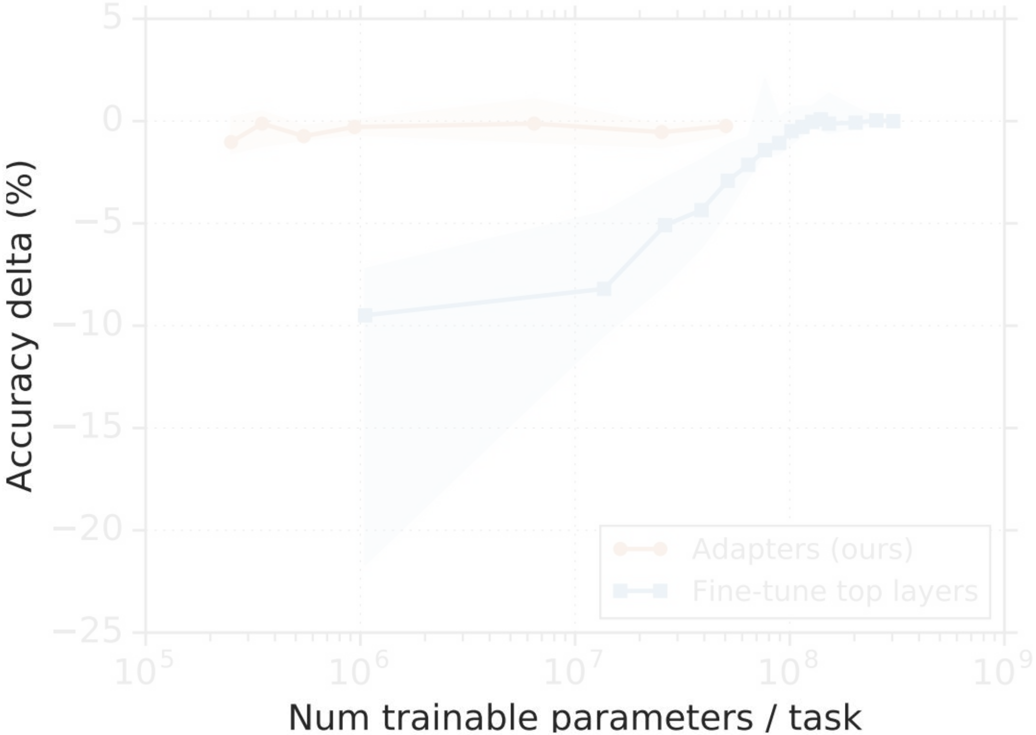
Results on GLUE Benchmark



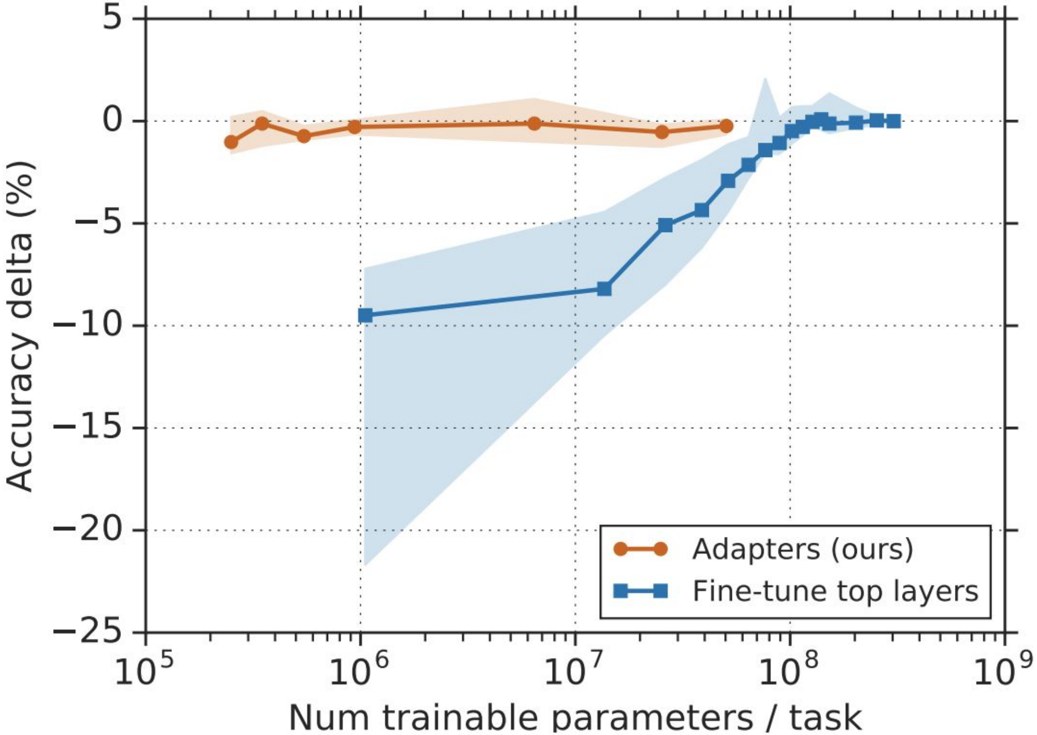
Results on GLUE Benchmark



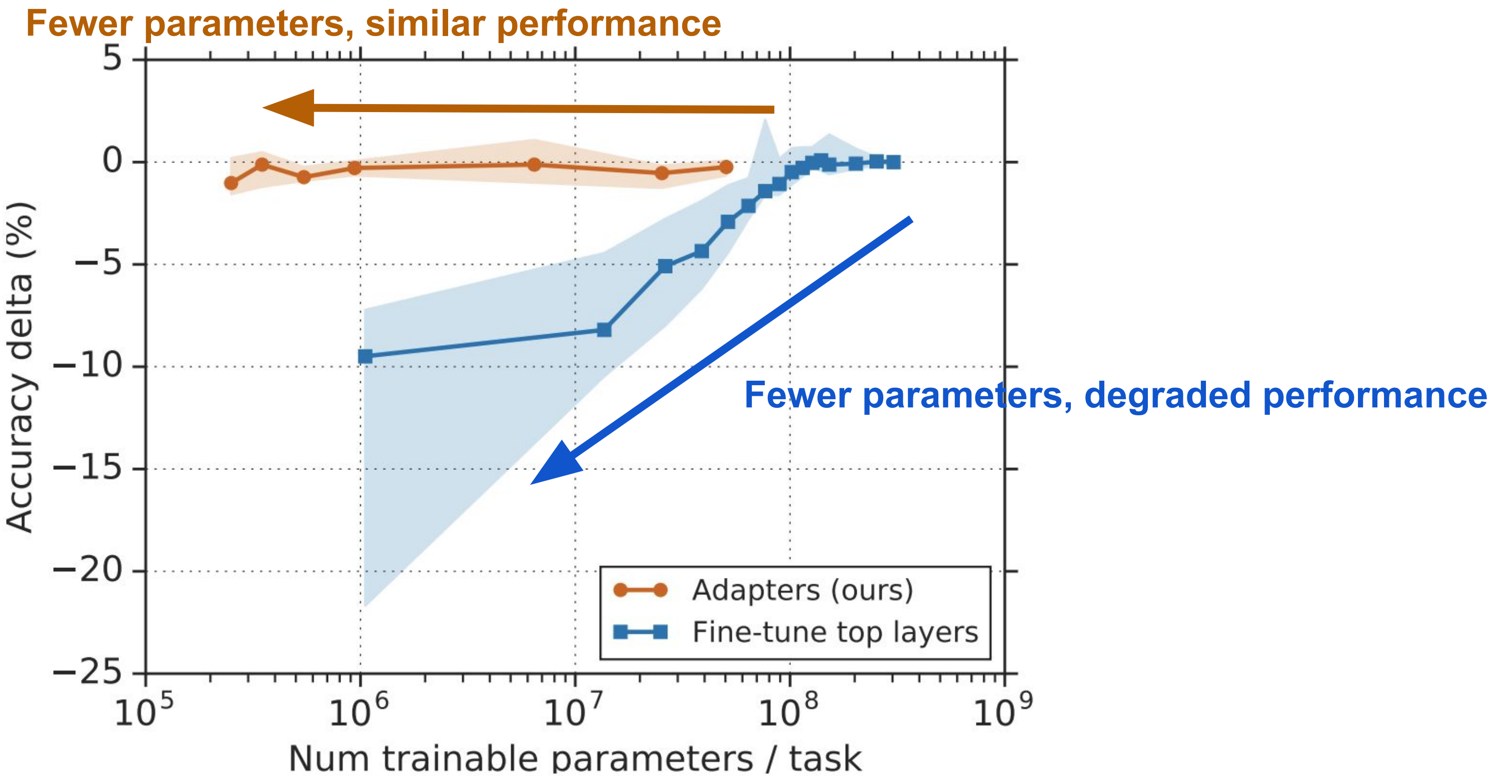
Results on GLUE Benchmark



Results on GLUE Benchmark



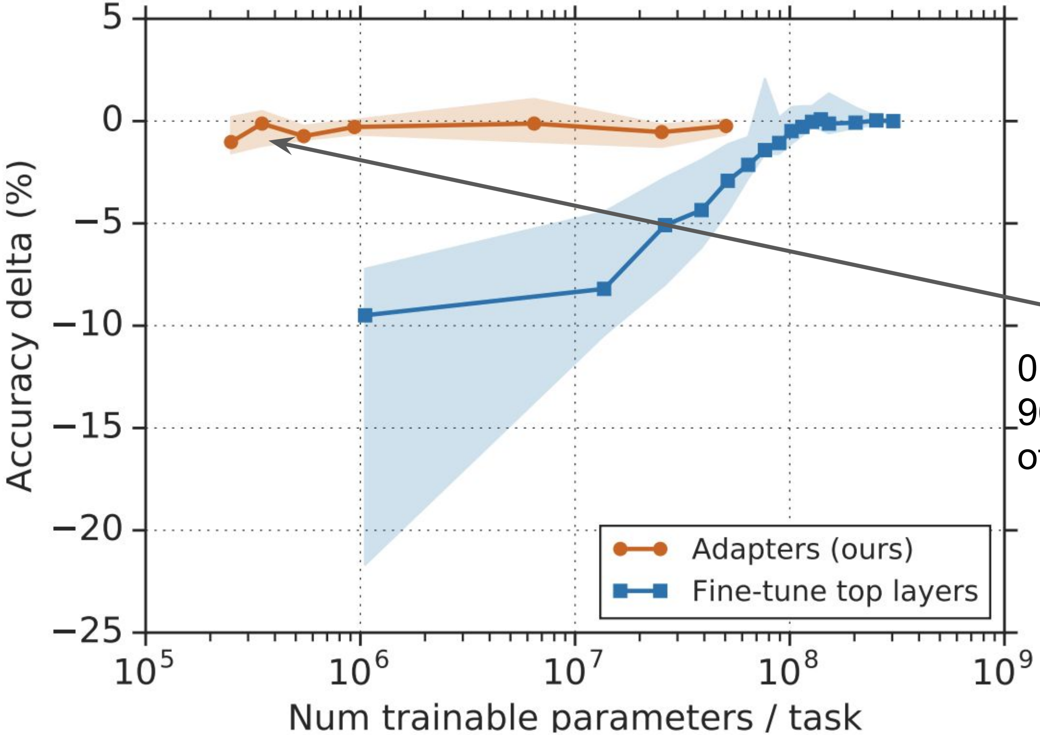
Results on GLUE Benchmark



Fewer parameters, similar performance

Fewer parameters, degraded performance

Results on GLUE Benchmark



0.4% accuracy drop for 96.4% reduction in the # of parameters/task

GREAT DEAL ★★★★★ GREAT DEAL ★★★★★ GREAT DEAL

Conclusions

1. If we move towards a single model future, **we need to improve parameter-efficiency of transfer learning**
2. We propose a **module reducing drastically # params/task for NLP**, e.g. by 30x at only 0.4% accuracy drop

Related work (@ ICML): ***"BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning", A. Stickland & I. Murray***

Please come to our poster today at 6:30 PM (#102)