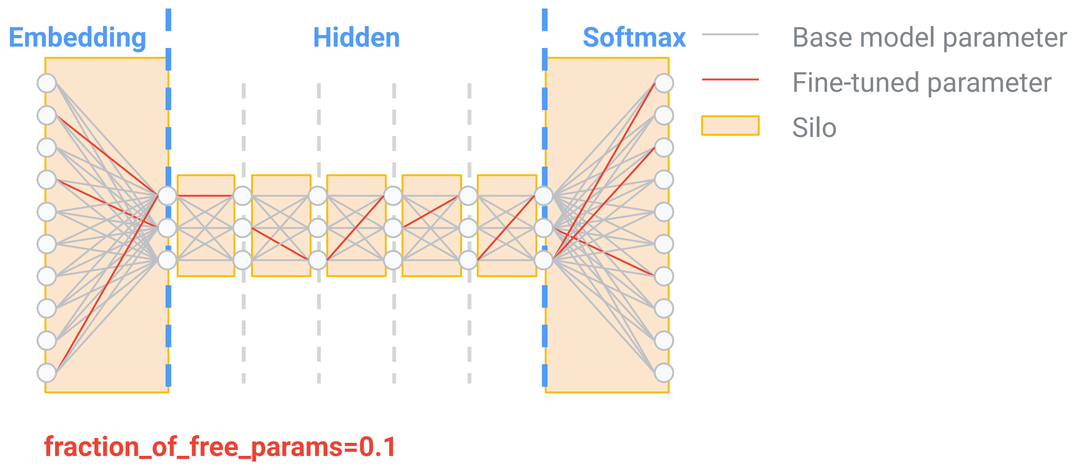
Google Research

Dynamic Subset Tuning: Expanding the Operational Range of Parameter-Efficient Training for Large Language Models

Felix Stahlberg, Jared Lichtarge, Shankar Kumar
{fstahlberg, lichtarge, shankarkumar}@google.com

Dynamic subset tuning

- Dynamic subset tuning (DST) is a **parameter-efficient training** approach that updates a small subset of the existing model parameters.
  - Strong and controllable regularization
  - More efficient storage
- DST is based on two core ideas:
  - **Siloing** encourages a more uniform distribution of the subset across the model.



  - **Dynamic subset selection** jointly optimizes the tunable parameters and the subset selection, i.e. the subset evolves during training.

Algorithm DST update function for computing the model parameters $\Theta^{(t+1)}$ for the next training iteration. Note that `compute_full_updates()` may have additional dependencies such as the optimizer state in momentum-based optimizers that we left out for the sake of simplicity.

Require: $\Theta^{(0)}$: Seed parameters at time step 0.
Require: $\Theta^{(t)}$: Parameters at time step t.
Require: ϵ: Fraction of free parameters.
Require: $\mathcal{S}$: Siloing partition.

1: $u \leftarrow$ `compute_full_updates`$(\Theta^{(t)})$ {Apply optimizer's update rule}
2: $\hat{\Theta} \leftarrow \Theta^{(t)} + u$
3: $\Delta \leftarrow d(\Theta^{(0)}, \hat{\Theta}, \Theta^{(0)})$ {Component-wise}
4: **for** $S \in \mathcal{S}$ **do**
5: $\quad q \leftarrow$ `quantile`$(\Delta\big|_S, 1-\epsilon)$
6: $\quad \Theta^{(t+1)}\big|_S \leftarrow$ `where`$(\Delta\big|_S > q, \hat{\Theta}\big|_S, \Theta^{(0)}\big|_S)$
7: **end for**
8: **return** $\Theta^{(t+1)}$

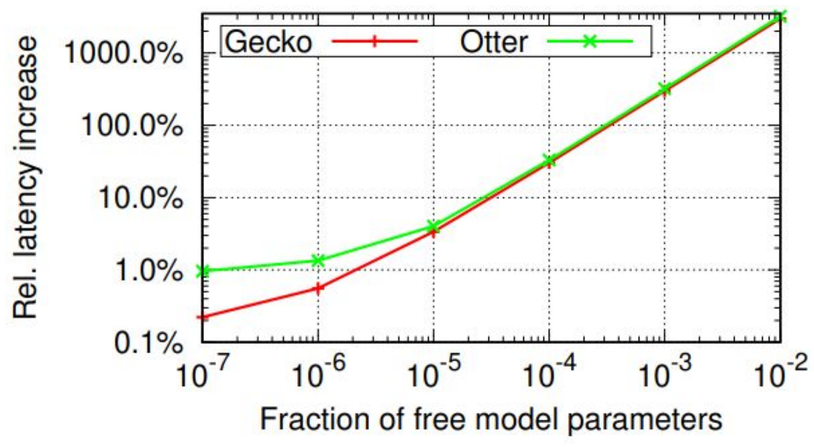
Compute full update

Compute component-wise distance to seed model

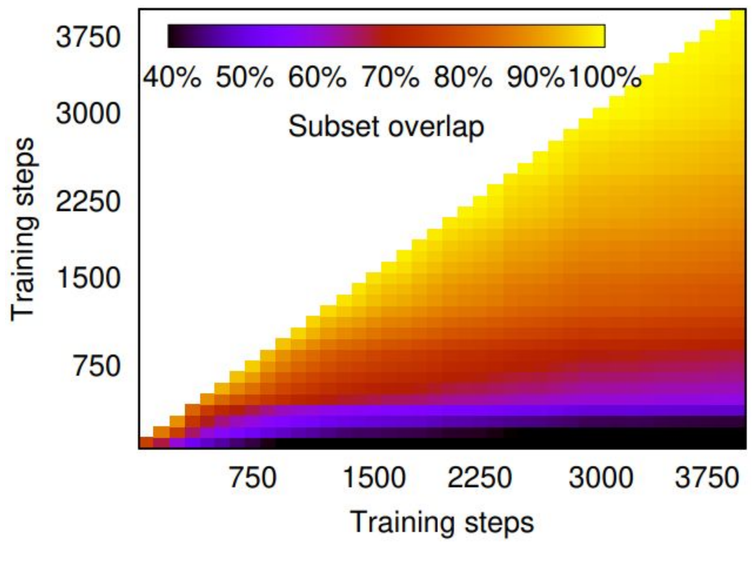
Enforce epsilon-constraint in each silo

Analyses

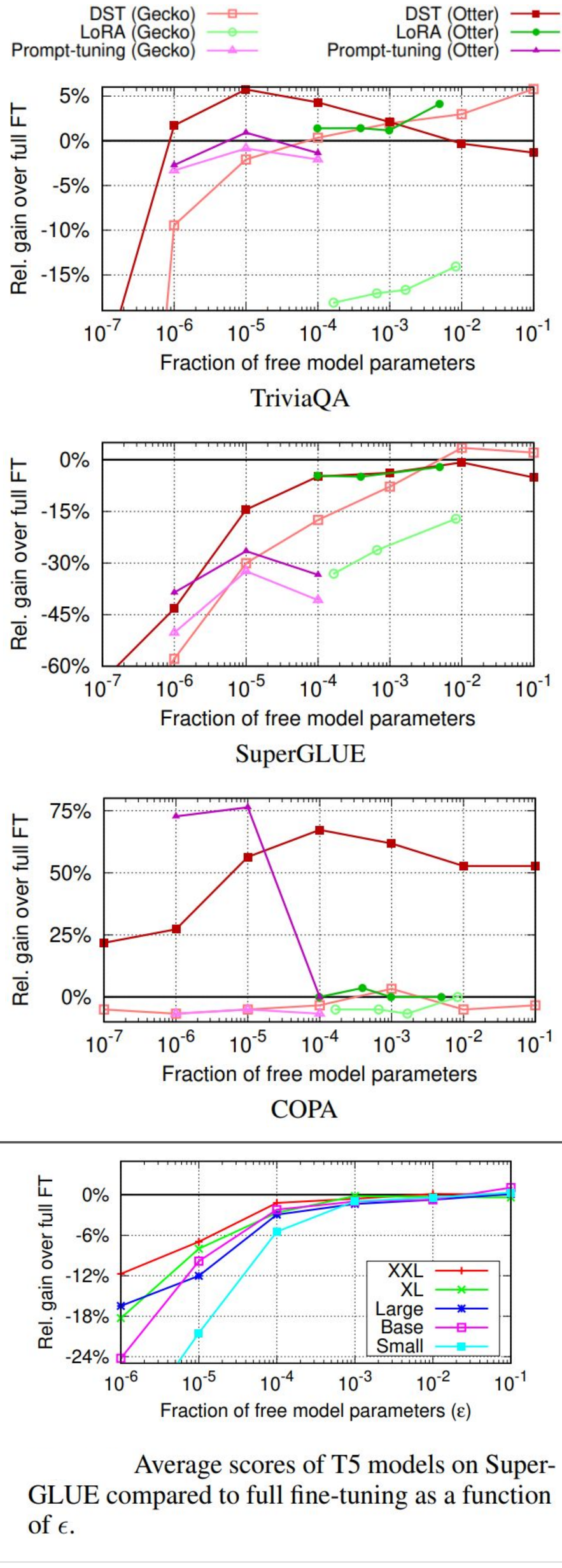
Serving latency (on-the-fly subset loading)



Subset convergence



Results



Average scores of T5 models on SuperGLUE compared to full fine-tuning as a function of ϵ.

Conclusion

- DST has a wider operational range than prompt-tuning and LoRA (down to ~1K free parameters for some tasks)
- Comparable or better performance to LoRA and prompt tuning with same parameter budget
- DST is suitable for fine-tuning large models on small datasets
- Small subsets can be loaded at inference time with small overhead