The Power of Scale for Parameter-Efficient Prompt Tuning

Brian Lester, Rami Al-Rfou, Noah Constant
{brianlester, rmyeid, nconstant}@google.com

Google Research

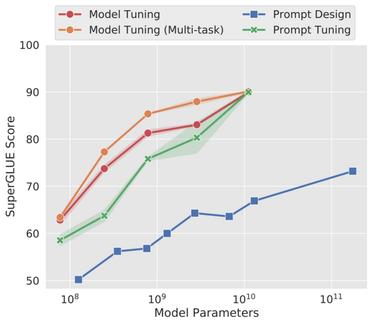
Prompt Tuning

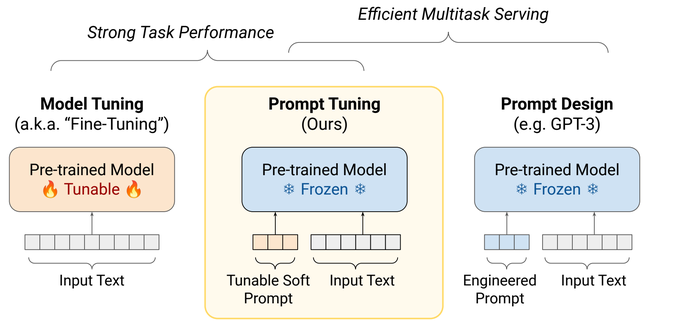
Prompt Tuning

- Keep pre-trained **T5 model frozen**.
- Learn **continuous prompts** for downstream tasks, prepended to input embeddings.

Closes Gap with Model Tuning

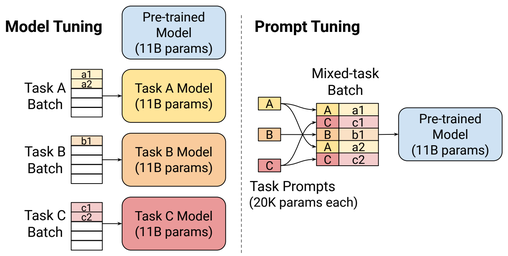
- As scale increases, matches model tuning.
- With <0.01% task-specific parameters!





Efficient Serving

- No model forking
- Tiny task-specification
- One model for all tasks
- Mixed-batch multitask inference



Ablation Study

Best Settings (— ✕ —)

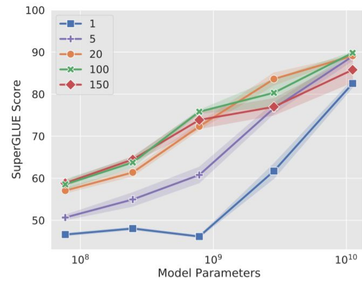
- **Prompt length**: 100
- **Prompt initialization**: Class Labels
- **Pre-training method**: LM Adaptation
- **LM adaptation steps**: 100K

Robustness with Scale

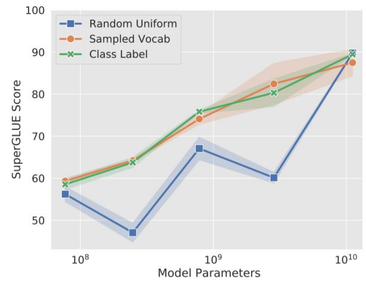
- For larger models, even non-optimal settings do well.

LM Adaptation

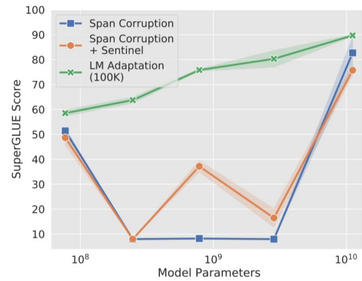
- T5 pre-training uses "sentinel" tokens, which are never seen in fine-tuning.
- Prepare model for prompting by training on standard LM task with realistic text.



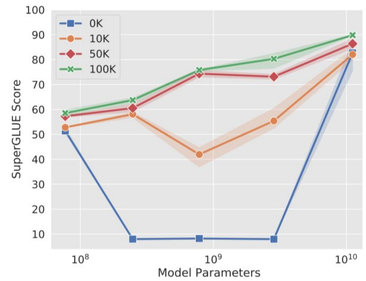
(a) Prompt length



(b) Prompt initialization



(c) Pre-training method



(d) LM adaptation steps

Parameter Efficiency & Resilience to Domain Shift

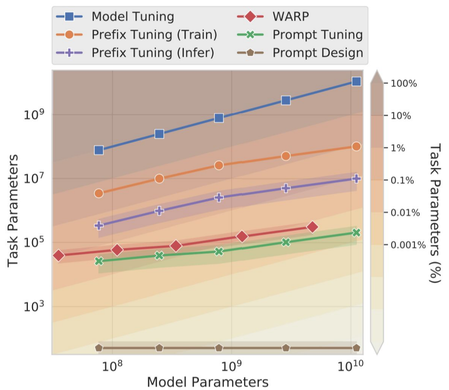
Parameter Efficiency

- For T5-XXL, we add just **0.003%** parameters!
- Parameters restricted to input layer (cf. "Prefix Tuning" per-layer).
- Adding ~3% parameters to T5-XXL (cf. "Adapters") ≈ adding BERT-Large.

Resilience to Domain Shift

- Train and checkpoint on SQuAD train/dev.
- Evaluate zero-shot on other QA tasks.
- Large gains, especially for larger domain shifts (TextbookQA: **+12.5**).

Dataset	Domain	Model	Prompt	Δ
SQuAD	Wiki	94.9 ±0.2	94.8 ±0.1	−0.1
TextbookQA	Book	54.3 ±3.7	**66.8** ±2.9	+12.5
BioASQ	Bio	77.9 ±0.4	**79.1** ±0.3	+1.2
RACE	Exam	59.8 ±0.6	**60.7** ±0.5	+0.9
RE	Wiki	88.4 ±0.1	**88.8** ±0.2	+0.4
DuoRC	Movie	**68.9** ±0.7	67.7 ±1.1	−1.2
DROP	Wiki	**68.9** ±1.7	67.1 ±1.9	−1.8



Prompt Ensembling & Interpretability

Prompt Ensembling

- Multi-task serving enables **efficient ensembling** by learning multiple prompts for one task.
- Prompt ensemble consistently beats best prompt in ensemble.

Dataset	Metric	Average	Best	Ensemble
BoolQ	acc.	91.1	91.3	**91.7**
CB	acc.	99.3	100.0	**100.0**
	F1	99.0	100.0	**100.0**
COPA	acc.	98.8	100.0	**100.0**
MultiRC	EM	65.7	66.3	**67.1**
	$F1_\alpha$	88.7	89.0	**89.4**
ReCoRD	EM/F1	92.7	92.9	**93.2**
	F1	99.4	93.5	**93.9**
RTE	acc.	92.6	**93.5**	**93.5**
WiC	acc.	76.2	76.6	**77.4**
WSC	acc.	95.8	**96.2**	**96.2**
SuperGLUE (dev)		90.5	91.0	**91.3**

Interpretability

- Examined top-5 nearest neighbor tokens by cosine similarity.
- Even with random init, **label verbalizers often appear** in top-5 nearest tokens.
- Nearest neighbors show strong "**word-like**" **clusters**:
  - Lexically-based clusters:
    - { *Technology / technology / Technologies / technological / technologies* }
  - Semantically-based clusters:
    - { *entirely / completely / totally / altogether / 100%* }

Selected References

[1] Li and Liang (2021) Prefix-Tuning: Optimizing Continuous Prompts for Generation
[2] Hambardzumyan et al. (2021) WARP: Word-level Adversarial ReProgramming
[3] Houlsby et al. (2019) Parameter-Efficient Transfer Learning for NLP