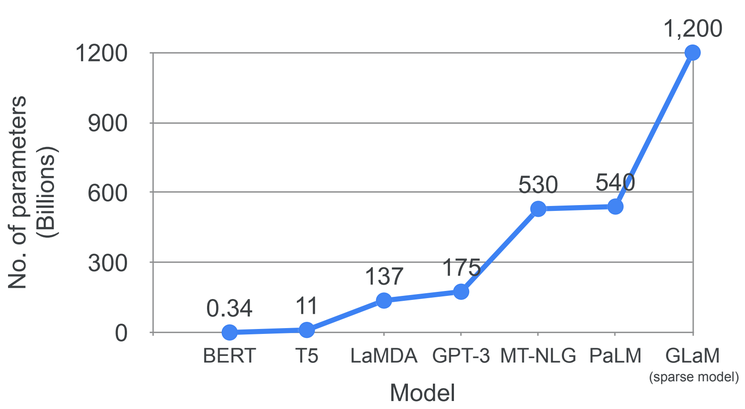
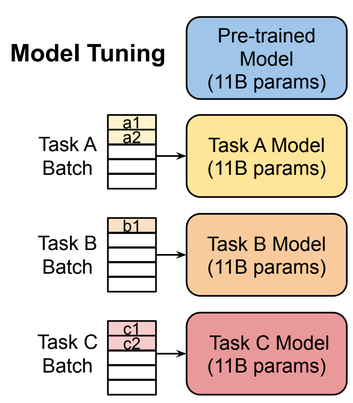
SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer

Tu Vu[1,2], Brian Lester[1], Noah Constant[1], Rami Al-Rfou[1], Daniel Cer[1]

Google Research[1] UMass Amherst[2]

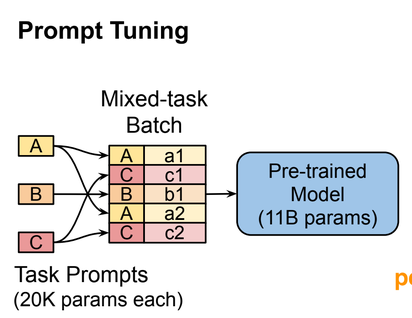
Pre-trained language models are costly to share and serve as model capacity increases



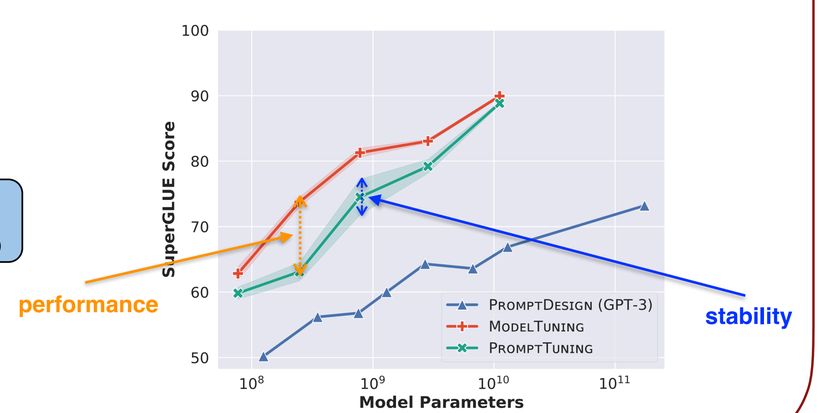


Lester et al., 2021

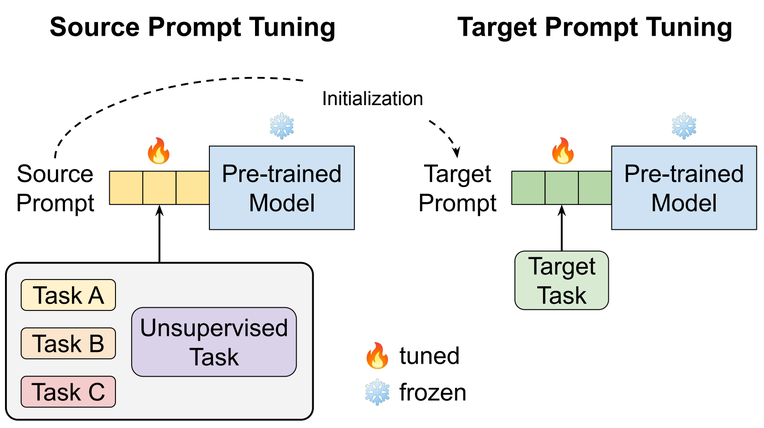
Prompt Tuning to the rescue... but there is still room for improvement!



Lester et al., 2021



Our generic SPoT approach



We learn a single generic source prompt on one or more source tasks, which is then used to initialize the prompt for each target task.

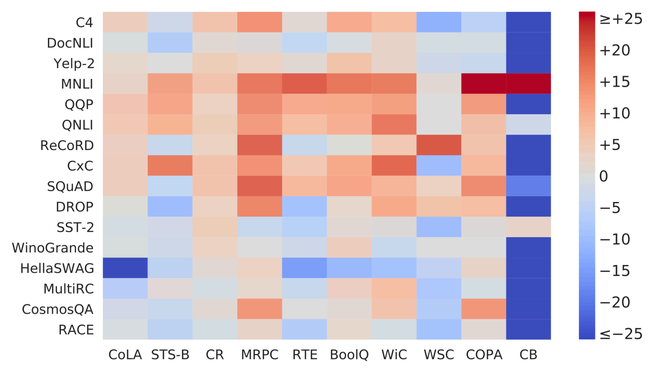
SPoT improves Prompt Tuning's performance & stability

Method	GLUE	SuperGLUE
Baseline		
PromptTuning	$81.2_{0.4}$	$66.6_{0.2}$
– longer tuning	$78.4_{1.7}$	$63.1_{1.1}$
SPoT *with different source mixtures*		
GLUE (*8 tasks*)	$\mathbf{82.8}_{0.2}$	$\mathbf{73.2}_{0.3}$
– longer tuning	$82.0_{0.2}$	$70.7_{0.4}$
C4	$82.0_{0.2}$	$67.7_{0.3}$
MNLI	$82.5_{0.0}$	$72.6_{0.8}$
SQuAD	$82.2_{0.1}$	$72.0_{0.4}$
SuperGLUE (8 tasks)	$82.0_{0.1}$	$66.6_{0.2}$
NLI (*7 tasks*)	$82.6_{0.1}$	$71.4_{0.2}$
Paraphrasing/similarity (*4 tasks*)	$82.2_{0.1}$	$69.7_{0.5}$
Sentiment (*5 tasks*)	$81.1_{0.2}$	$68.6_{0.1}$
MRQA (*6 tasks*)	$81.8_{0.2}$	$68.4_{0.2}$
RAINBOW (*6 tasks*)	$80.3_{0.6}$	$64.0_{0.4}$
Translation (*3 tasks*)	$82.4_{0.2}$	$65.3_{0.1}$
Summarization (*9 tasks*)	$80.9_{0.3}$	$67.1_{1.0}$
GEM (*8 tasks*)	$81.9_{0.2}$	$70.5_{0.5}$
All (C4 + 55 supervised tasks)	$81.8_{0.2}$	$67.9_{0.9}$

SPoT helps close the gap with Model Tuning across model sizes

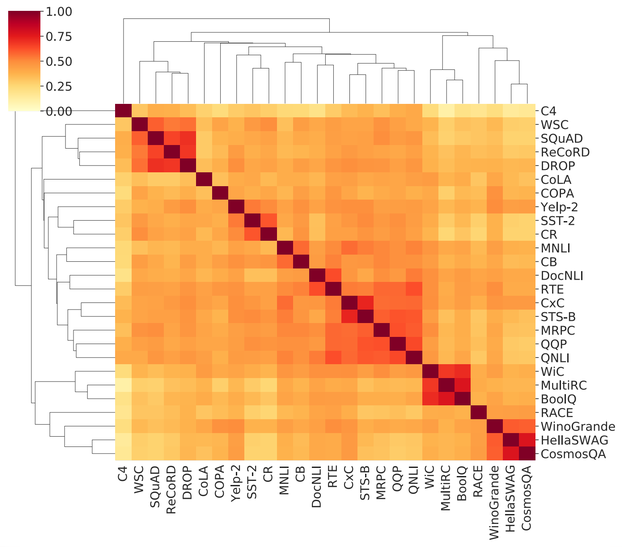
With a score of 89.2 on SuperGLUE, SPoT is the first parameter-efficient approach that is competitive with methods that tune billions of parameters.

Many tasks benefit each other via prompt transfer

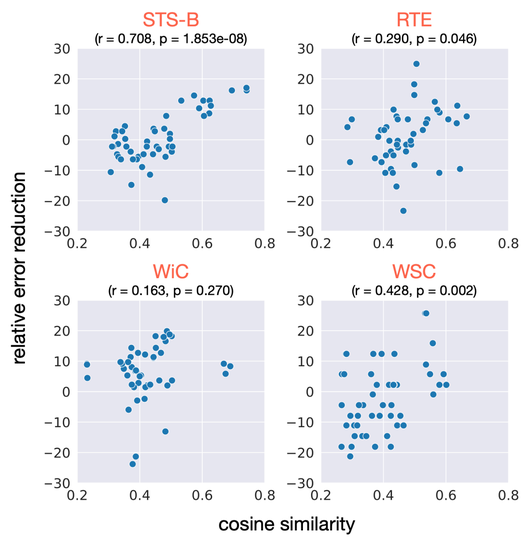


A large-scale study on prompt transferability with 26 NLP tasks (16 source tasks, 10 target tasks, 160 source-target combinations).

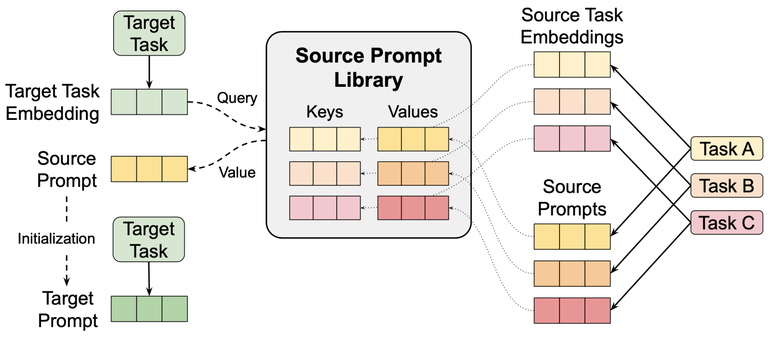
Task prompts capture task relationships



Correlation between task prompt similarity & task transferability



Our targeted SPoT approach



We learn separate prompts for various source tasks, saving early checkpoints as task embeddings and best checkpoints as source prompts. These form the keys and values of our prompt library. Given a novel target task, a user: (i) computes a task embedding, (ii) retrieves an optimal source prompt, and (iii) trains a target prompt, initialized from the source prompt.

Retrieving targeted source tasks via task embeddings is helpful

Method	Change		Avg. score
	Abs.	Rel.	
Baseline	-	-	$74.7_{0.7}$
Brute-force Search ($k = 48$)			
Oracle	$6.0_{0.5}$	$26.5_{1.1}$	$80.7_{0.0}$
Cosine Similarity of Average Tokens			
Best of Top-k			
$k = 1$	$1.5_{0.5}$	$11.7_{1.1}$	$76.2_{0.1}$
$k = 3$	$2.7_{0.6}$	$16.6_{1.1}$	$77.4_{0.3}$
$k = 6$	$3.8_{0.1}$	$20.0_{1.1}$	$78.5_{0.5}$
$k = 9$	$4.5_{0.4}$	$22.2_{1.1}$	$79.2_{0.1}$
$k = 12$	$5.0_{0.9}$	$23.6_{2.2}$	$79.7_{0.4}$
$k = 15$	$5.4_{0.8}$	$24.9_{1.8}$	$80.1_{0.3}$
Per-token Average Cosine Similarity			
Best of Top-k			
$k = 1$	$2.0_{0.4}$	$12.1_{1.1}$	$76.7_{0.7}$
$k = 3$	$2.9_{0.6}$	$17.0_{0.6}$	$77.5_{0.4}$
$k = 6$	$4.5_{0.5}$	$22.1_{1.2}$	$79.2_{0.1}$
$k = 9$	$4.6_{0.5}$	$22.6_{0.9}$	$79.5_{0.2}$
$k = 12$	$5.0_{0.6}$	$23.5_{1.4}$	$79.6_{0.1}$
$k = 15$	$5.3_{0.9}$	$24.5_{2.2}$	$80.0_{0.4}$
Top-k Weighted Average			
best $k = 3$	$1.9_{0.5}$	$11.5_{2.7}$	$76.6_{0.1}$
Top-k Multi-task Mixture			
best $k = 12$	$3.1_{0.5}$	$15.3_{2.8}$	$77.8_{0.1}$

Task embeddings provide an effective means of predicting and exploiting task transferability, eliminating 69% of the source task search space while keeping 90% of the best-case quality gain.

Conclusion

- ✦ We show that scale is not necessary for Prompt Tuning to match Model Tuning's performance; SPOT matches or beats Model Tuning across all model sizes.
- ✦ We conduct a large-scale and systematic study on task transferability in the context of prompt tuning.
- ✦ We propose an efficient retrieval method that measures task embedding similarity to identify which tasks could benefit each other.
- ✦ Our library of task prompts, pre-trained models, and practical recommendations are available at https://github.com/google-research/prompt-tuning/tree/main/prompt_tuning/spot.

References

✦ Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In EMNLP 2021, pages 3045–3059.