SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer

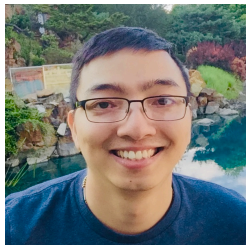
Tu Vu
Google Research
University of Massachusetts Amherst

May, 2022

Google

SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer

Paper: go/spot
go/soft-prompt-transfer



Tu Vu

Brian Lester

Noah Constant

Rami Al-Rfou

Daniel Cer

Google AI

Agenda

Prompt Tuning

Improving Prompt Tuning through Soft Prompt Transfer

Predicting and Exploiting Transferability between Tasks via Task Prompts

Conclusion & Future Work

Q & A

Agenda

Prompt Tuning

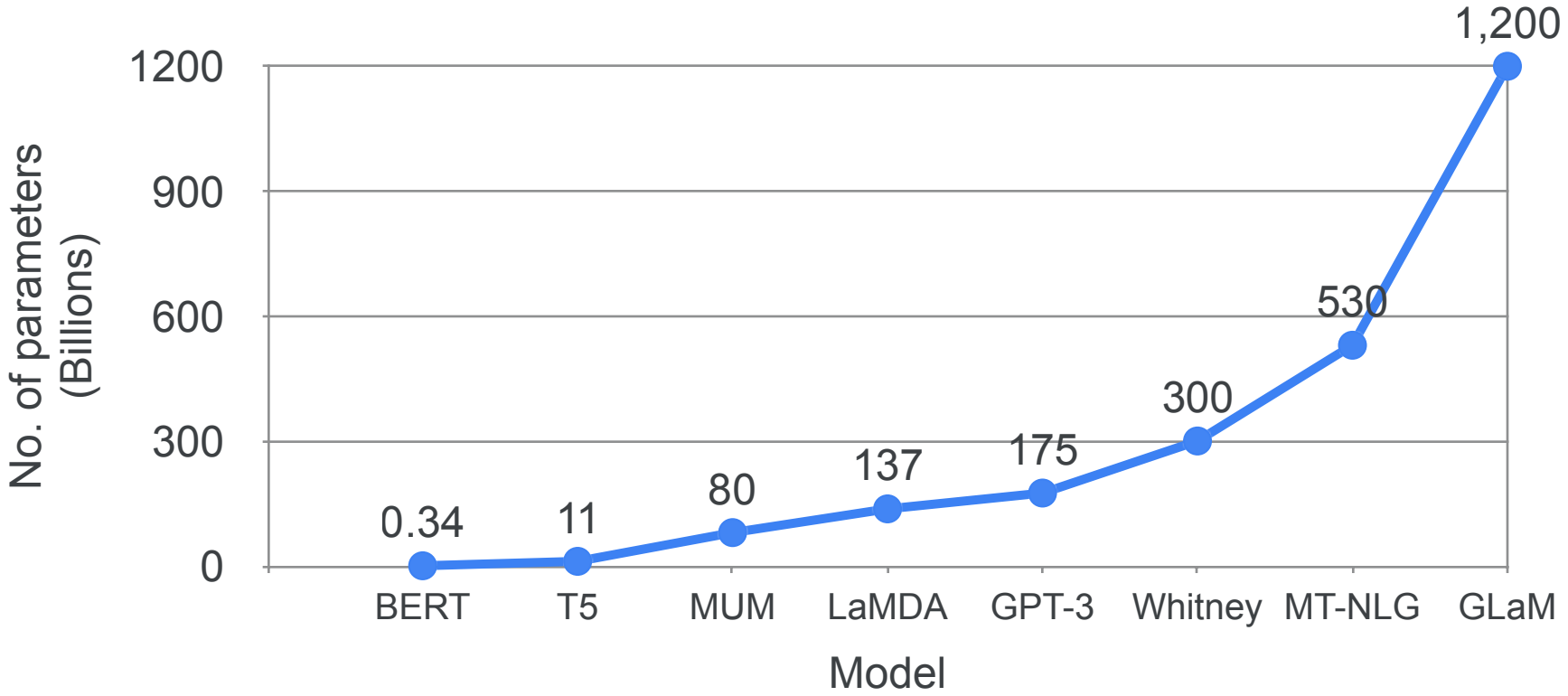
Improving Prompt Tuning through Soft Prompt Transfer

Predicting and Exploiting Transferability between Tasks via Task Prompts

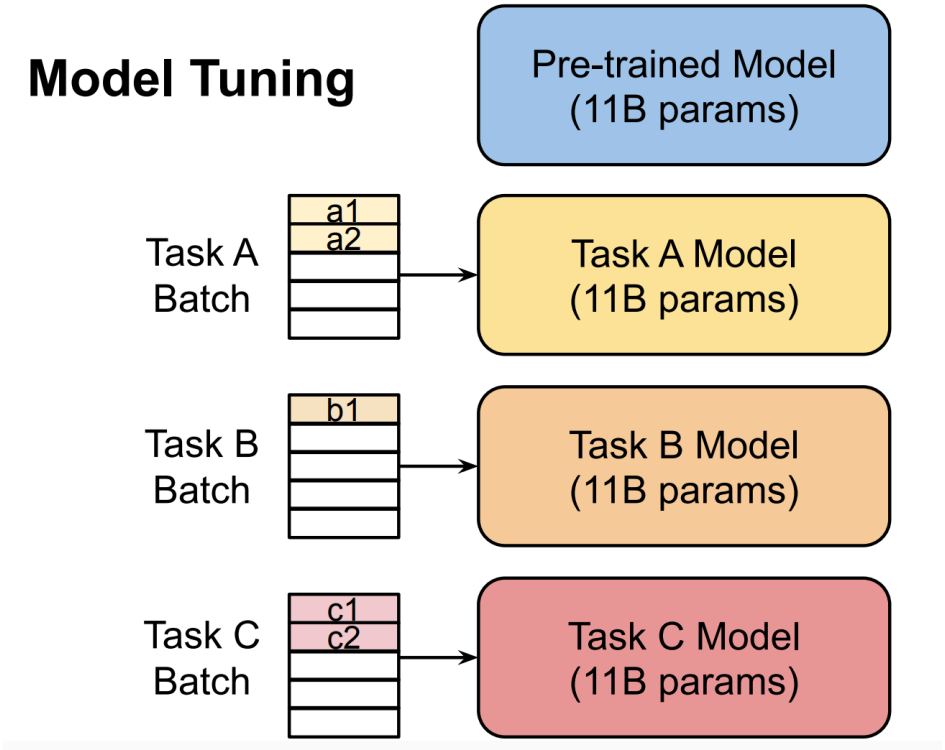
Conclusion & Future Work

Q & A

Scaling up the model size has continued to push the boundaries of possibility

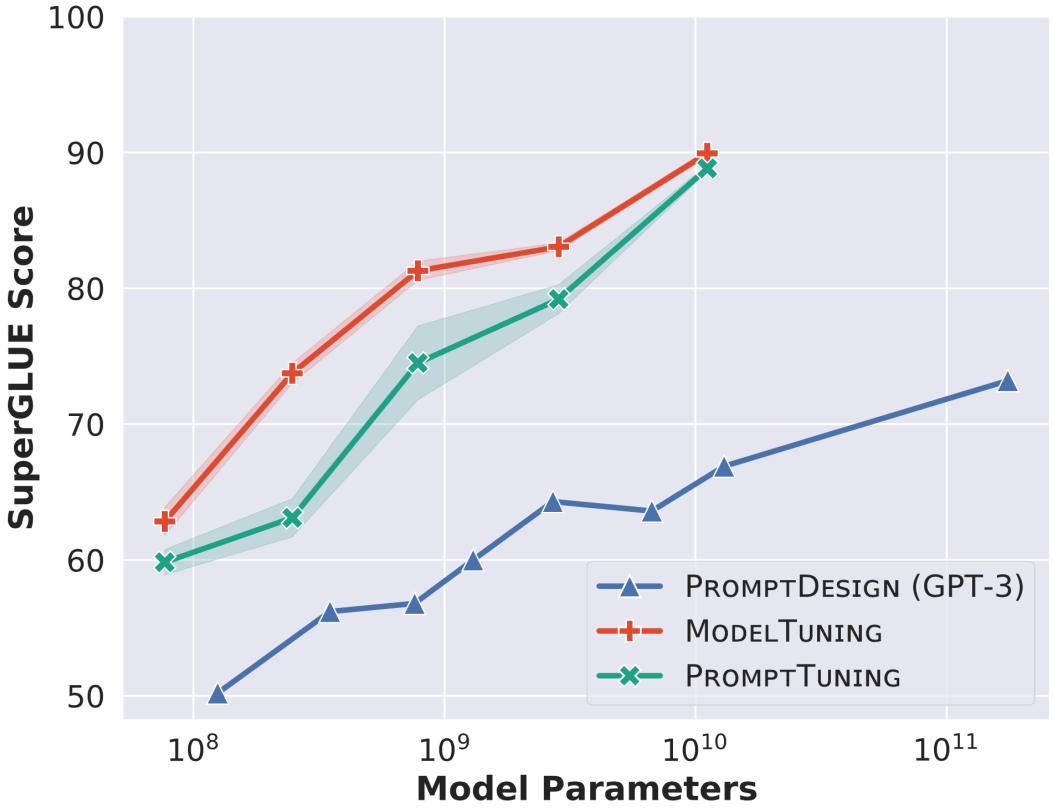


Practical challenges: large-scale models are costly to share and serve

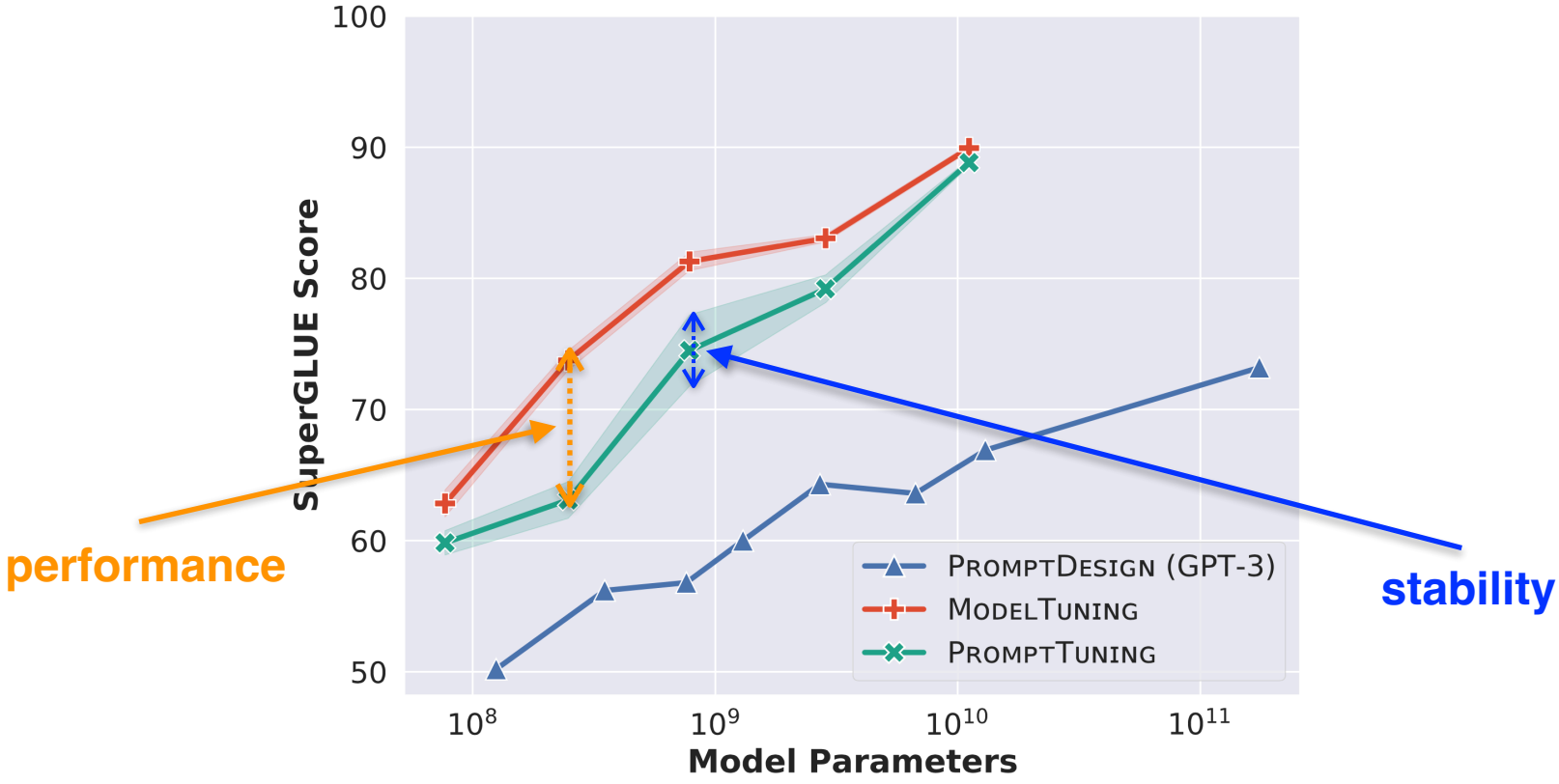


Lester et al., 2021

Prompt Tuning becomes competitive with Model Tuning as model capacity increases



Room for improving Prompt Tuning



Agenda

Prompt-based learning & Prompt Tuning

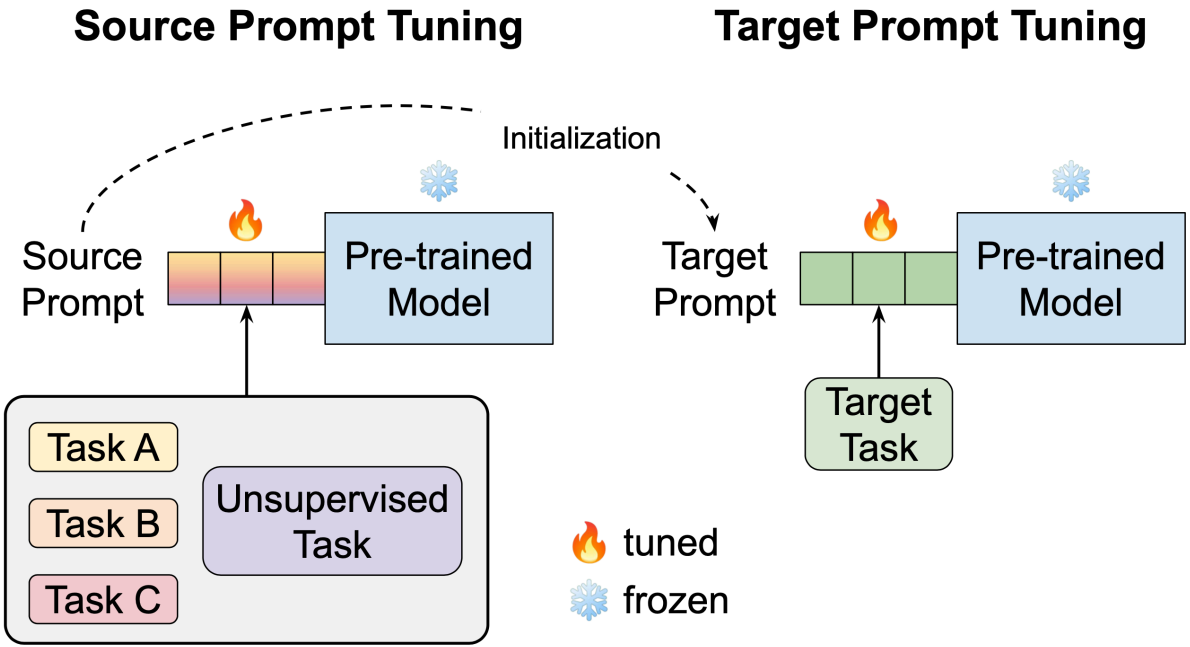
Improving Prompt Tuning through Soft Prompt Transfer

Predicting and Exploiting Transferability between Tasks via Task Prompts

Conclusion & Future Work

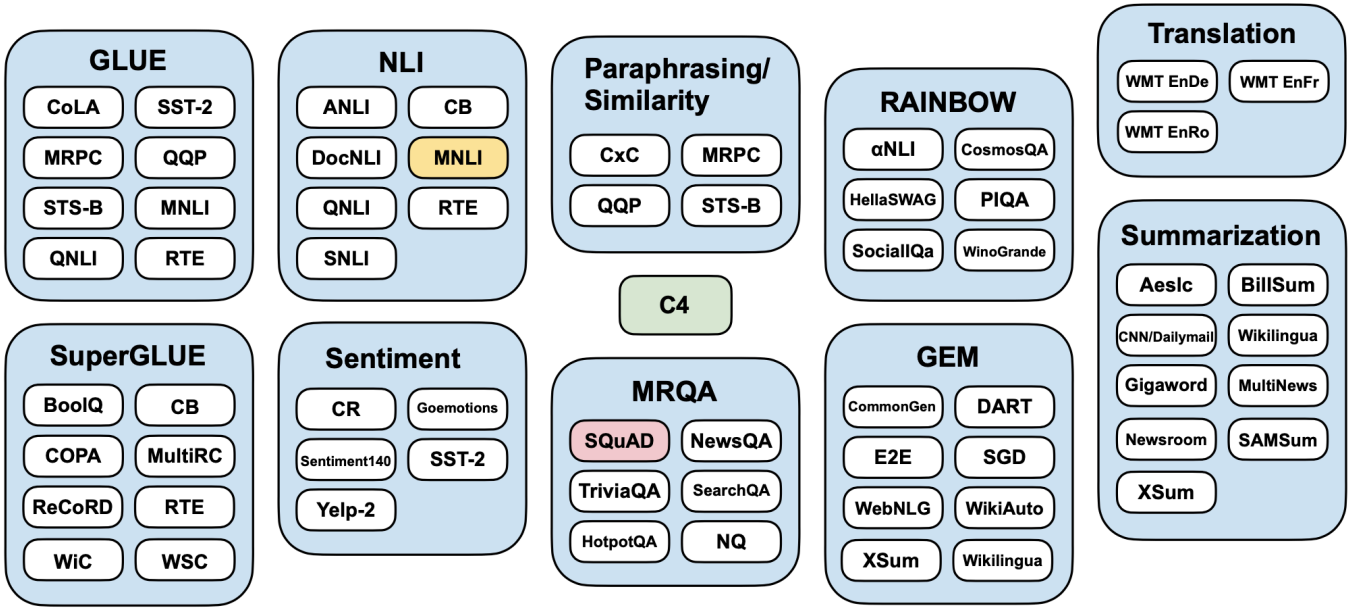
Q & A

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.

Mixing datasets from different NLP benchmarks / task families



Datasets used in our experiments. C4, MNLI, and SQuAD were all used by themselves as single source tasks in addition to being mixed in with other tasks.

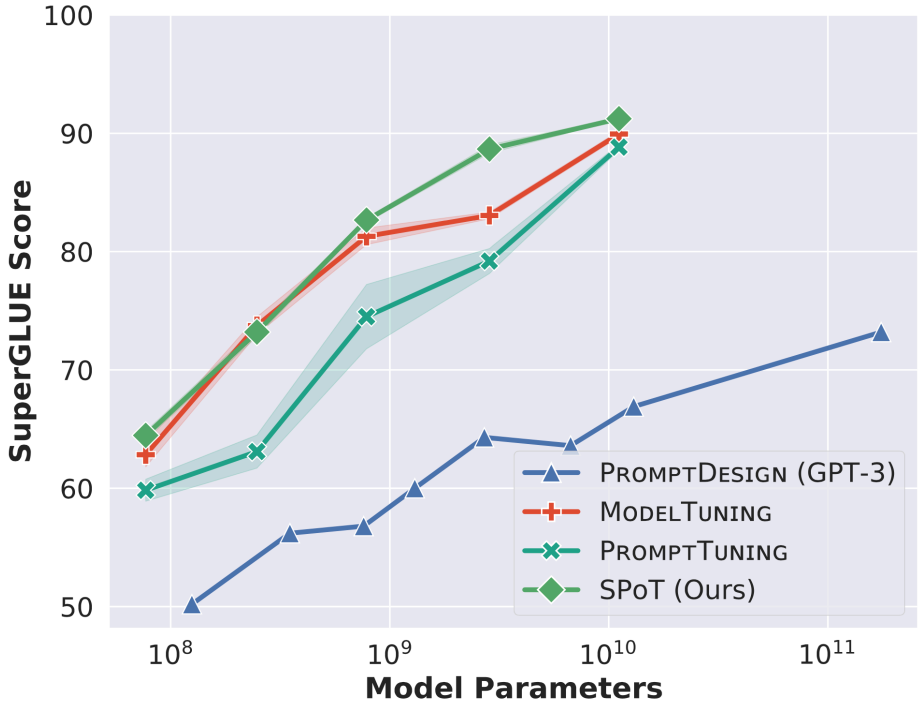
SPoT significantly improves performance and stability of Prompt Tuning

GLUE and SUPERGLUE results achieved by applying T5 BASE with different prompt tuning approaches. We report the mean and standard deviation (in the subscript) across three random seeds.

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)	$81.8_{0.2}$	$67.9_{0.9}$

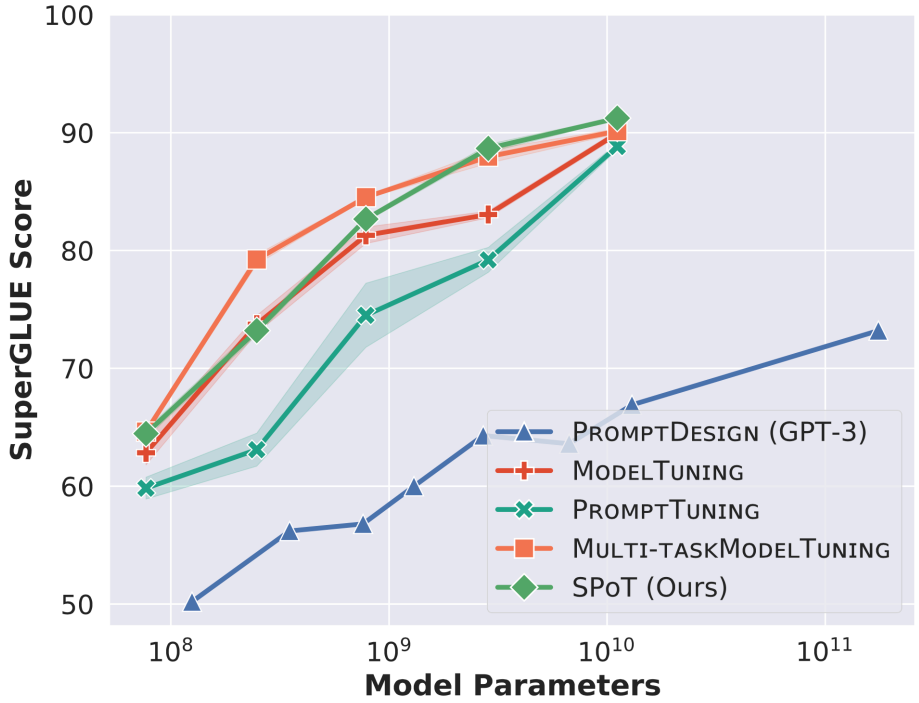
SPoT helps close the gap with Model Tuning across model sizes

Our SPoT approach—which transfers a prompt learned from a mixture of source tasks (here, GLUE) onto target tasks—outperforms vanilla PROMTTUNING and GPT-3 on SUPERGLUE by a large margin, matching or outperforming MODELTUNING across all model sizes.



An apples-to-apples comparison to Multi-task Model Tuning

At the XXL model size, SPoT even outperforms MULTI-TASKMODELTUNING, which fine-tunes the entire model on the GLUE mixture before fine-tuning it individually on each SUPERGLUE task.



Agenda

Prompt-based learning & Prompt Tuning

Improving Prompt Tuning through Soft Prompt Transfer

Predicting and Exploiting Transferability between Tasks via Task Prompts

Conclusion & Future Work

Q & A

A large-scale study on task transferability in the context of prompt tuning

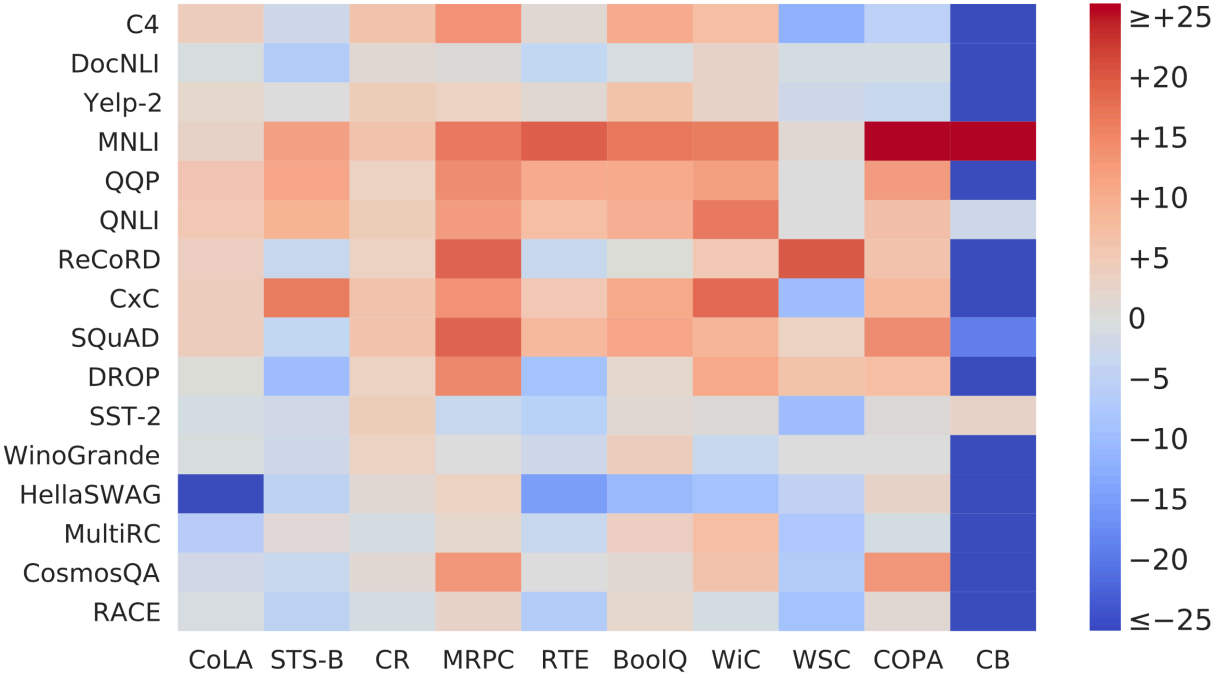
26 NLP tasks

- 16 source tasks, 10 target tasks, 160 source-target combinations of tasks
- covering various task types

Name	Task type	|Train|
16 source tasks		
C4	language modeling	365M
DocNLI	NLI	942K
Yelp-2	sentiment analysis	560K
MNLI	NLI	393K
QQP	paraphrase detection	364K
QNLI	NLI	105K
ReCoRD	QA	101K
CxC	semantic similarity	88K
SQuAD	QA	88K
DROP	QA	77K
SST-2	sentiment analysis	67K
WinoGrande	commonsense reasoning	40K
HellaSWAG	commonsense reasoning	40K
MultiRC	QA	27K
CosmosQA	commonsense reasoning	25K
RACE	QA	25K
10 target tasks		
BoolQ	QA	9K
CoLA	grammatical acceptability	9K
STS-B	semantic similarity	6K
WiC	word sense disambiguation	5K
CR	sentiment analysis	4K
MRPC	paraphrase detection	4K
RTE	NLI	2K
WSC	coreference resolution	554
COPA	QA	400
CB	NLI	250

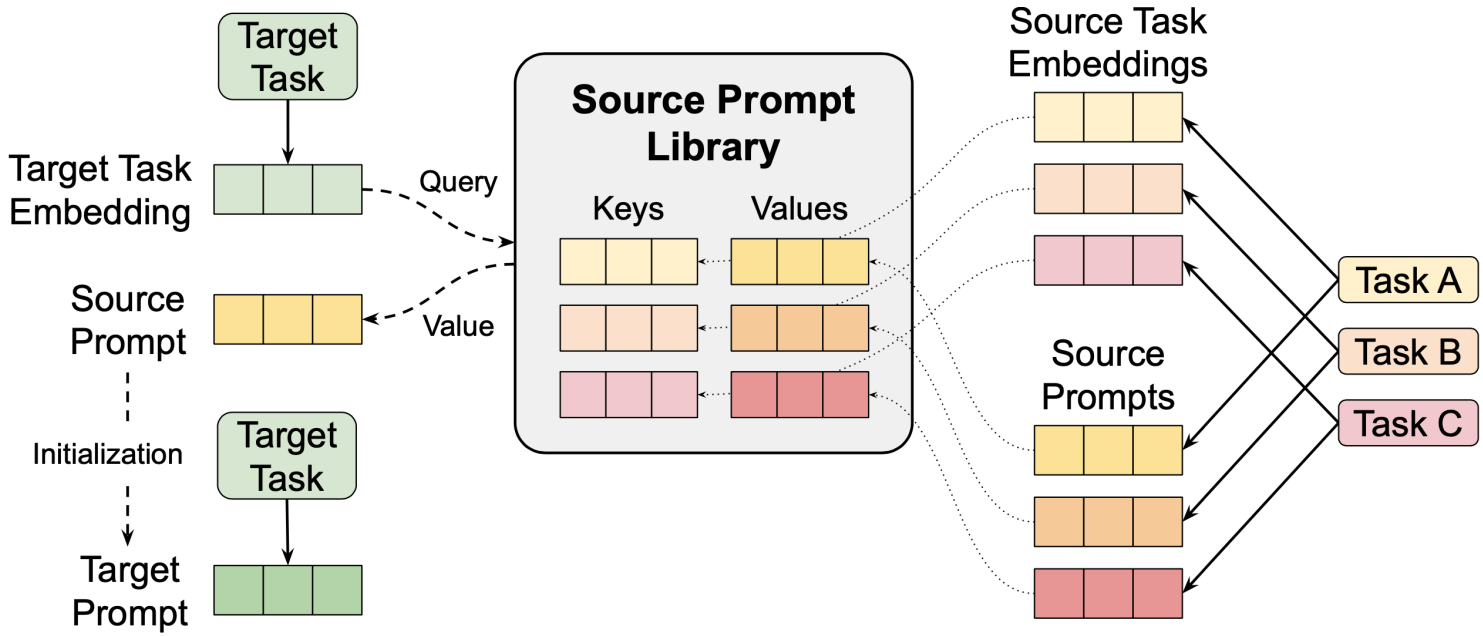
Tasks used in our task transferability experiments, sorted by training dataset size.

Many tasks can benefit each other via prompt transfer



A heatmap of our task transferability results. Each cell shows the relative error reduction on the target task of the transferred prompt from the associated source task (row) to the associated target task (column).

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

Measuring task similarity through prompts

Cosine Similarity of Average Tokens

- cosine similarity between the average pooled representations of the prompt tokens:

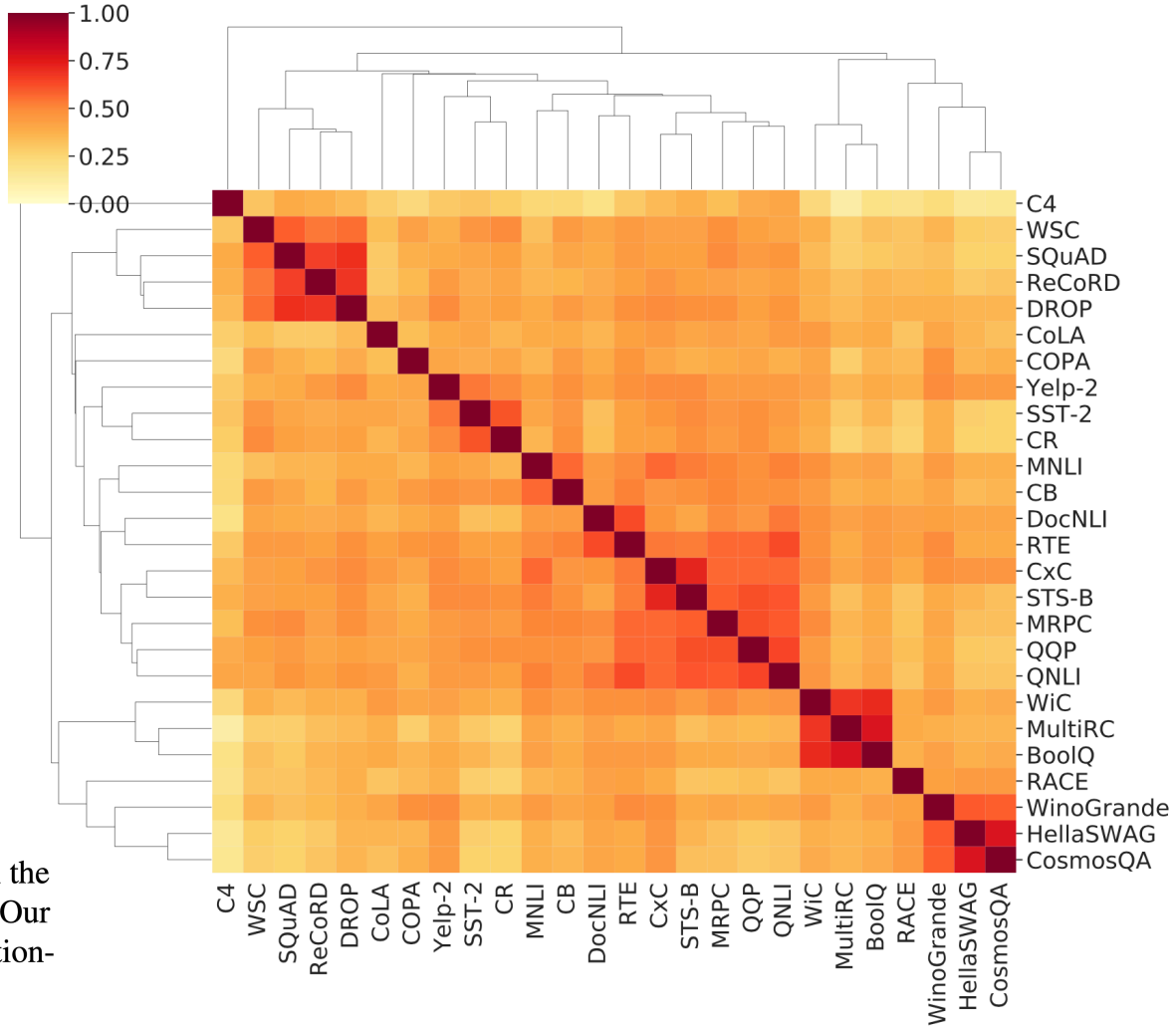
$$sim(t^1, t^2) = cos(\frac{1}{\mathcal{L}} \sum_i \boldsymbol{e}_i^1, \frac{1}{\mathcal{L}} \sum_j \boldsymbol{e}_j^2)$$

Per-token Average Cosine Similarity

- average cosine similarity between every prompt token pair

$$sim(t^1, t^2) = \frac{1}{\mathcal{L}^2} \sum_i \sum_j cos(\boldsymbol{e}_i^1, \boldsymbol{e}_j^2)$$

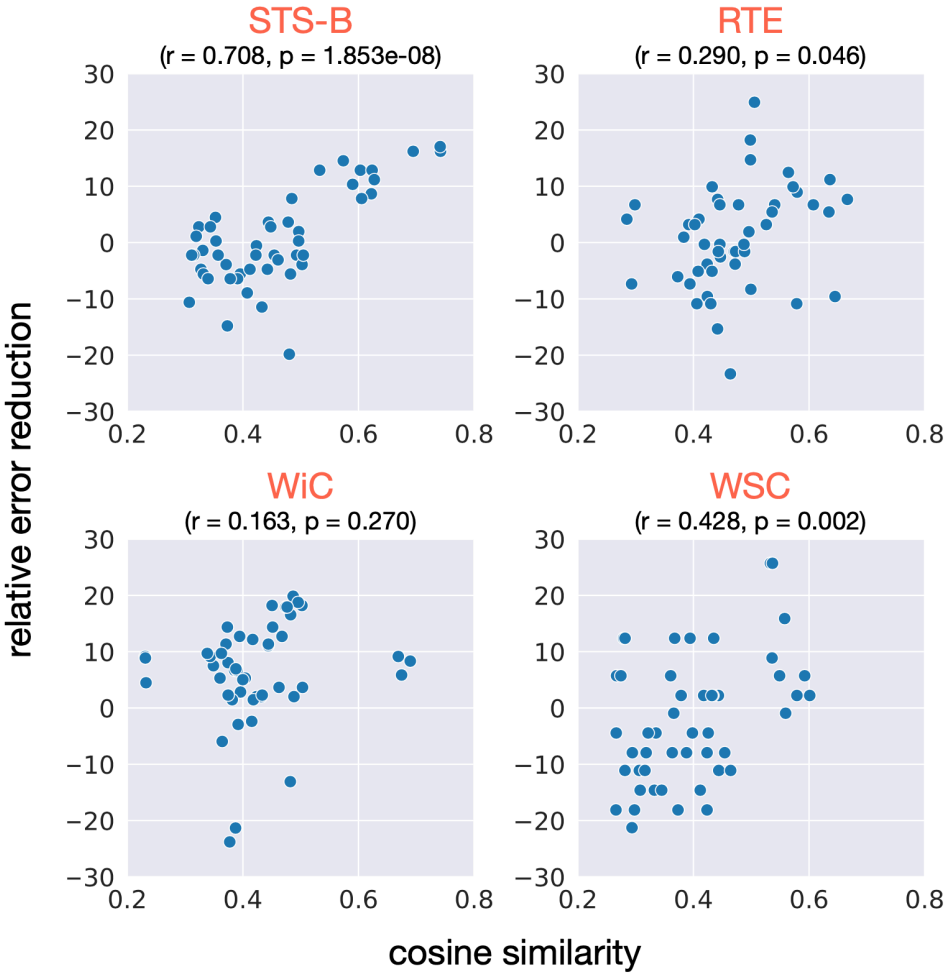
Task embeddings capture task relationships



A clustered heatmap of cosine similarities between the task embeddings of the 26 NLP tasks we study. Our prompt-based task embeddings capture task relationships: similar tasks cluster together.

Correlation between task similarity & task transferability

Correlation between task similarity and task transferability. Each point represents a source prompt. The x-axis shows the cosine similarity between the associated source and target task embeddings, averaged over three runs for the target task (orange title). The y-axis measures the relative error reduction on the target task achieved by each source prompt. We include the Pearson correlation coefficient (r) and p-value.



Predicting transferability via similarity

Best of Top-k

- select the top-k source prompts and use each of them individually for the target prompt; this requires prompt tuning k times on the target task

Top-k Weighted Average

- initialize the target prompt with a weighted average of the top-k source prompts so that we only perform prompt tuning on the target task once

Top-k Multi-task Mixture

- mix source datasets whose prompts are in the top-k prompts and the target dataset together, and then perform source prompt tuning on this multi-task mixture

Retrieving targeted source tasks via task embeddings is helpful

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 obtained by oracle selection.

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 $k = 12$	$3.1_{0.5}$	$15.3_{2.8}$	$77.8_{0.1}$

Agenda

Prompt-based learning & Prompt Tuning

Improving Prompt Tuning through Soft Prompt Transfer

Predicting and Exploiting Transferability between Tasks via Task Prompts

Conclusion & Future Work

Q & A

Conclusion

- We show that scale is not necessary for Prompt Tuning to match the performance of Model Tuning; our proposed SPOT approach 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.
- We will release our library of task prompts and pre-trained models, and provide practical recommendations for adapting our library to NLP practitioners.

Future work

- Prompt-based Cross-lingual transfer
  - soft prompts as language/task representations
  - identify the most beneficial source languages/tasks for a given novel target task in a novel target language

Agenda

Prompt-based learning & Prompt Tuning

Improving Prompt Tuning through Soft Prompt Transfer

Predicting and Exploiting Transferability between Tasks via Task Prompts

Conclusion & Future Work

Q & A

Thank you!