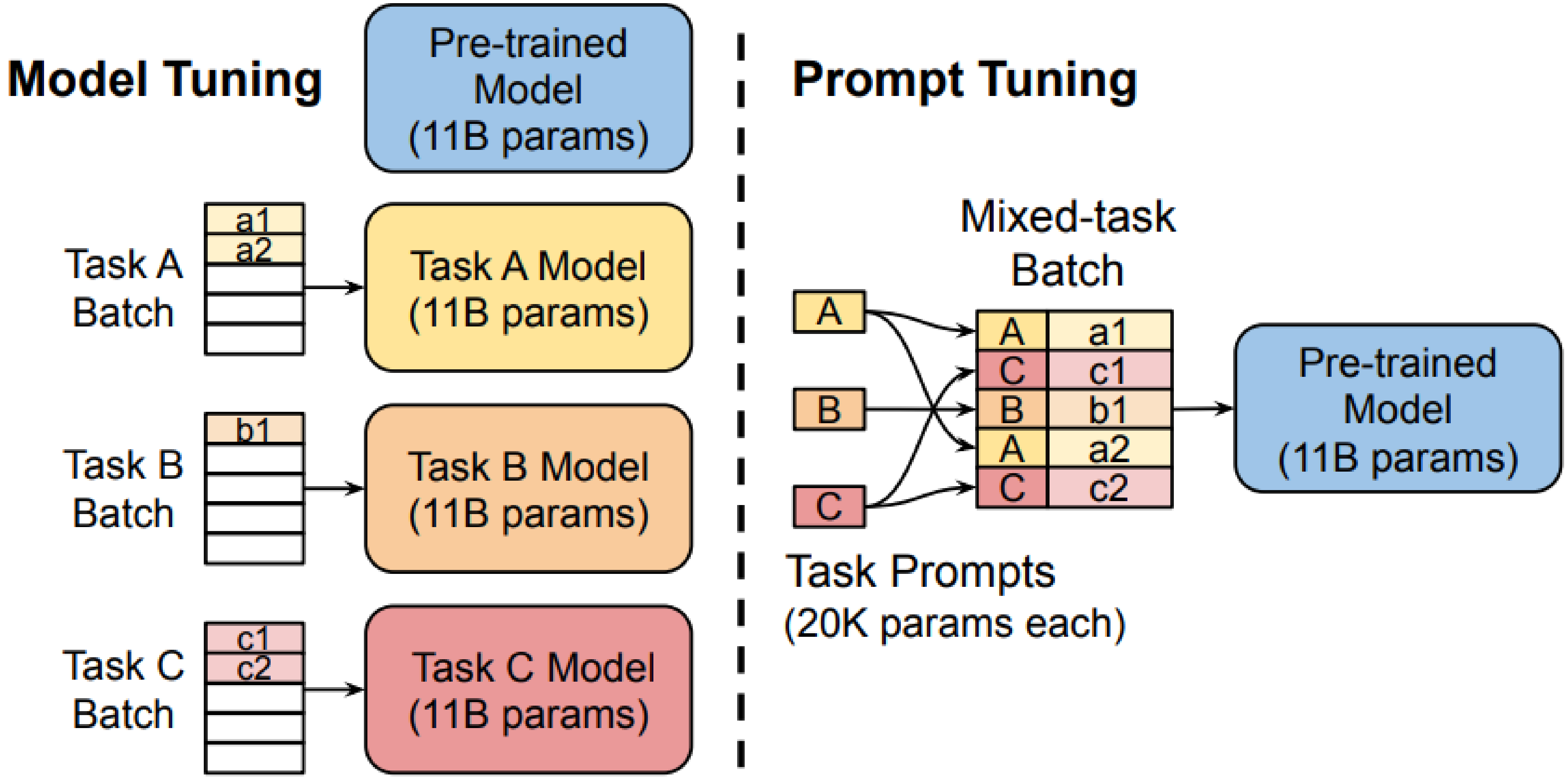
Prompt-tuning: сравнение методов

Гвоздева Дарья БПМИ211

План на сегодня

- Идея prompt-tuning’a
- Parameter-Efficient Prompt Tuning
  - Суть подхода
  - Результаты эксперимента
  - Сравнение с прочими подходами
- Pre-trained Prompt Tuning
  - Суть подхода
  - Эффективность
- Итоги


Model Tuning
Pre-trained Model (11B params)
Task A Batch
a1
a2
Task A Model (11B params)
Task B Batch
b1
Task B Model (11B params)
Task C Batch
c1
c2
Task C Model (11B params)
Prompt Tuning
Mixed-task Batch
A
B
C
A a1
C c1
B b1
A a2
C c2
Pre-trained Model (11B params)
Task Prompts (20K params each)

Основная идея

$$\Pr(y|X) -> Pr_{theta,theta_p}(Y|[P;X])$$

Основная идея

$$\mathrm{Pr}(y|X) -> Pr_{theta,theta_p}(Y|[P;X])$$

GPT3: P= $\{p_1, \ldots, p_n\}$ - p_i меняются, theta fixed

PT: P (prompt)- $fixed$, theta_p fixed & own for each prompt, embeddings updated

Prompt design: обновляемые p_i из фиксированных эмбеддингов

Подходы к дизайну инициализации

Представление промпта

- с нуля случайно
- каждый токен - эмбеддинг из словаря модели
- классификация: эмбеддинги выходных классов

Результаты

Setup:

- T5 различных размеров
- Дефолтная конфигурация в 100 токенов
- Benchmark: SuperGLUE
  - Каждый промпт – своя задача
  - Без мультитаска/смешивания данных
- 30.000 steps
- T5 -> стандартное отклонение + кросс энтропия
- lr: const 0.3
- Batchsize: 32
- Optimizer^ Adafactor

Baselines:
(1) T5 model tuning, default
(2) T5 multitask

+ сравнение с GPT-3 few-shot

Результаты

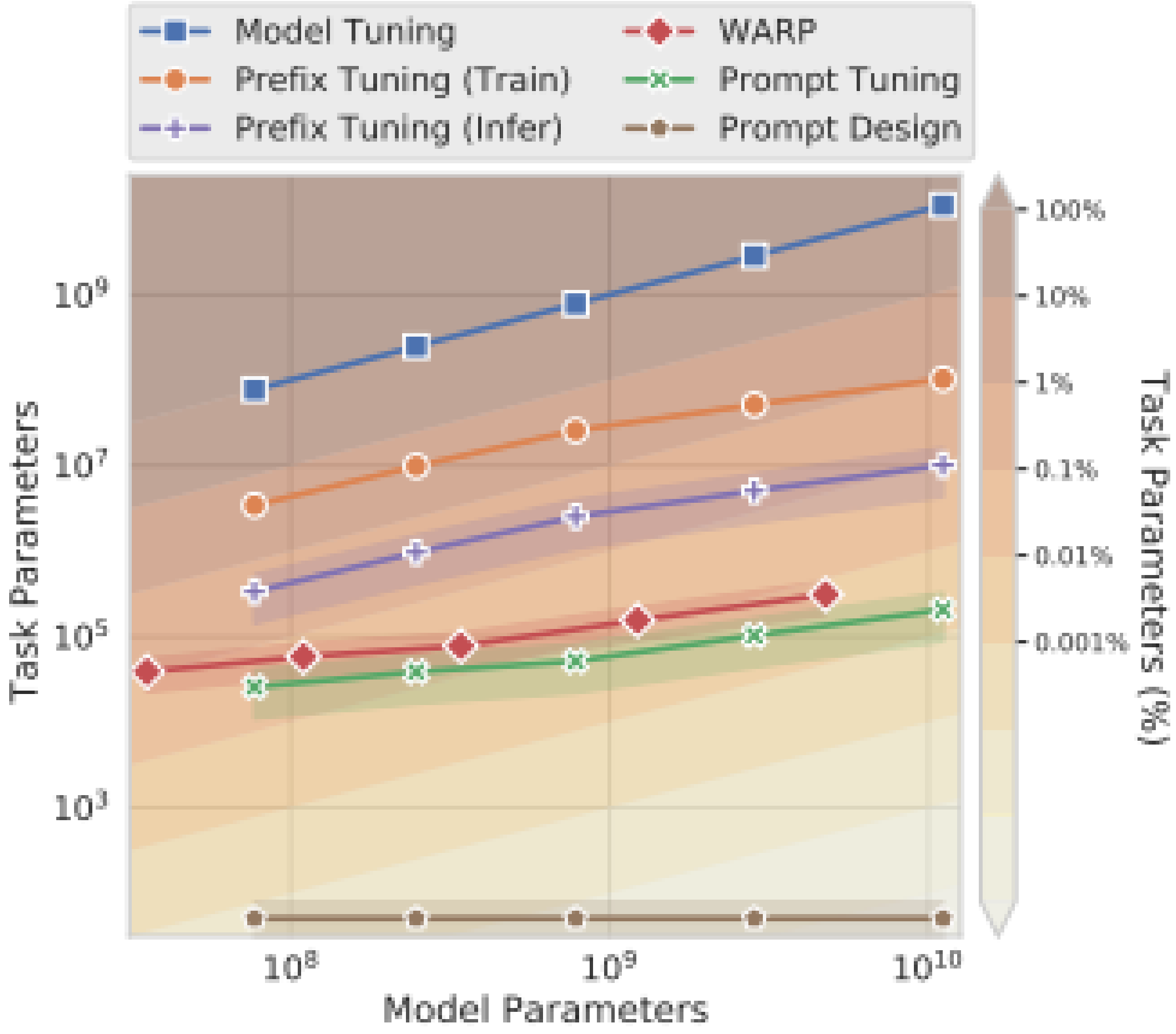
Model Tuning: все параметры специфичны для задания

Prefix Tuning: активации в prefix каждого слоя, 0.1–1% task-specific параметров для задачи, но больше для обучения

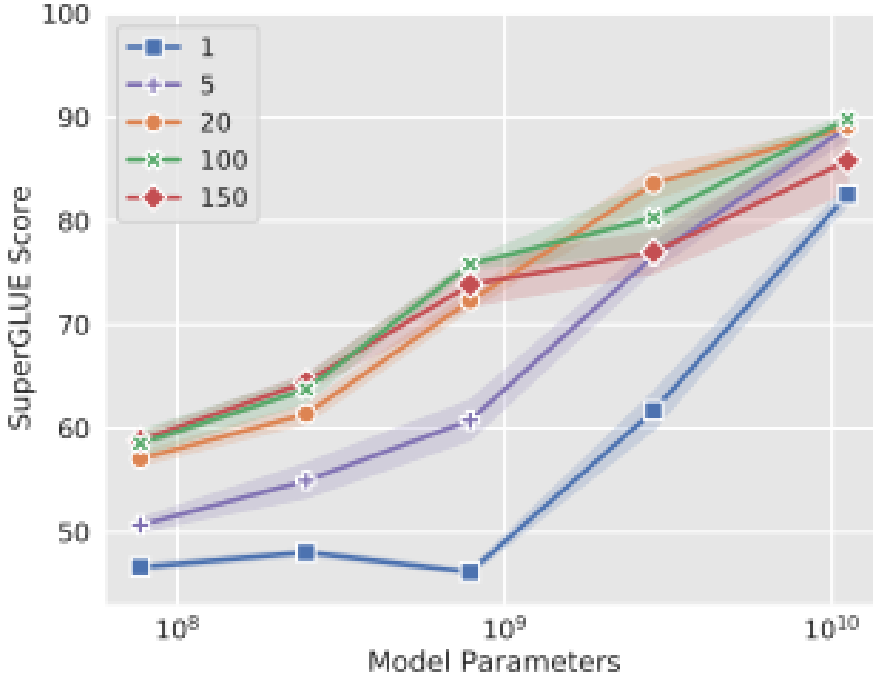
WARP: 0.1%, tuning входных/выходных слоев

Prompt Tuning: только prompt embeddings

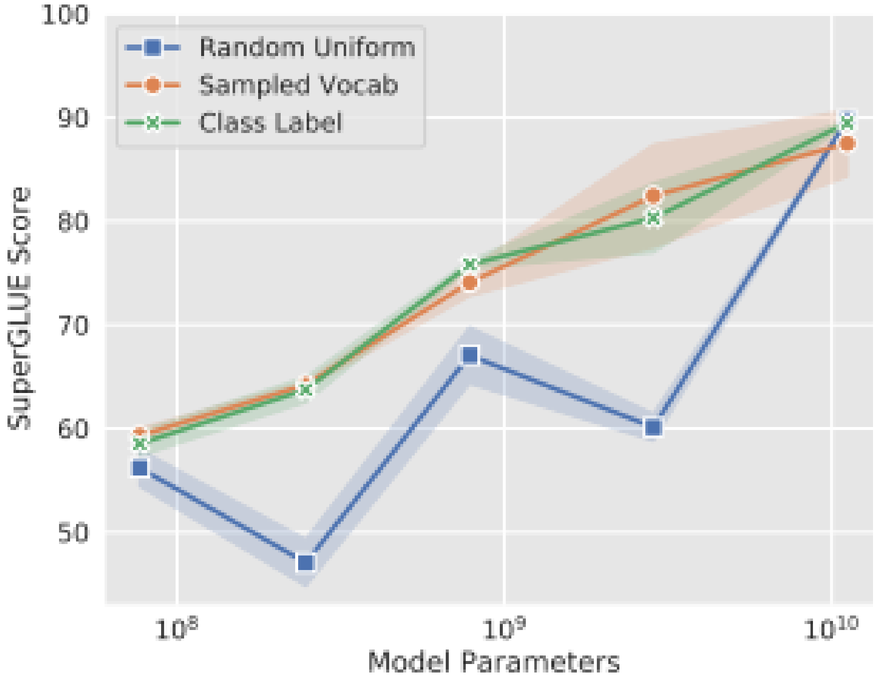
Prompt Design: только последовательность prompt IDs (500–2000 токенов)



Результаты

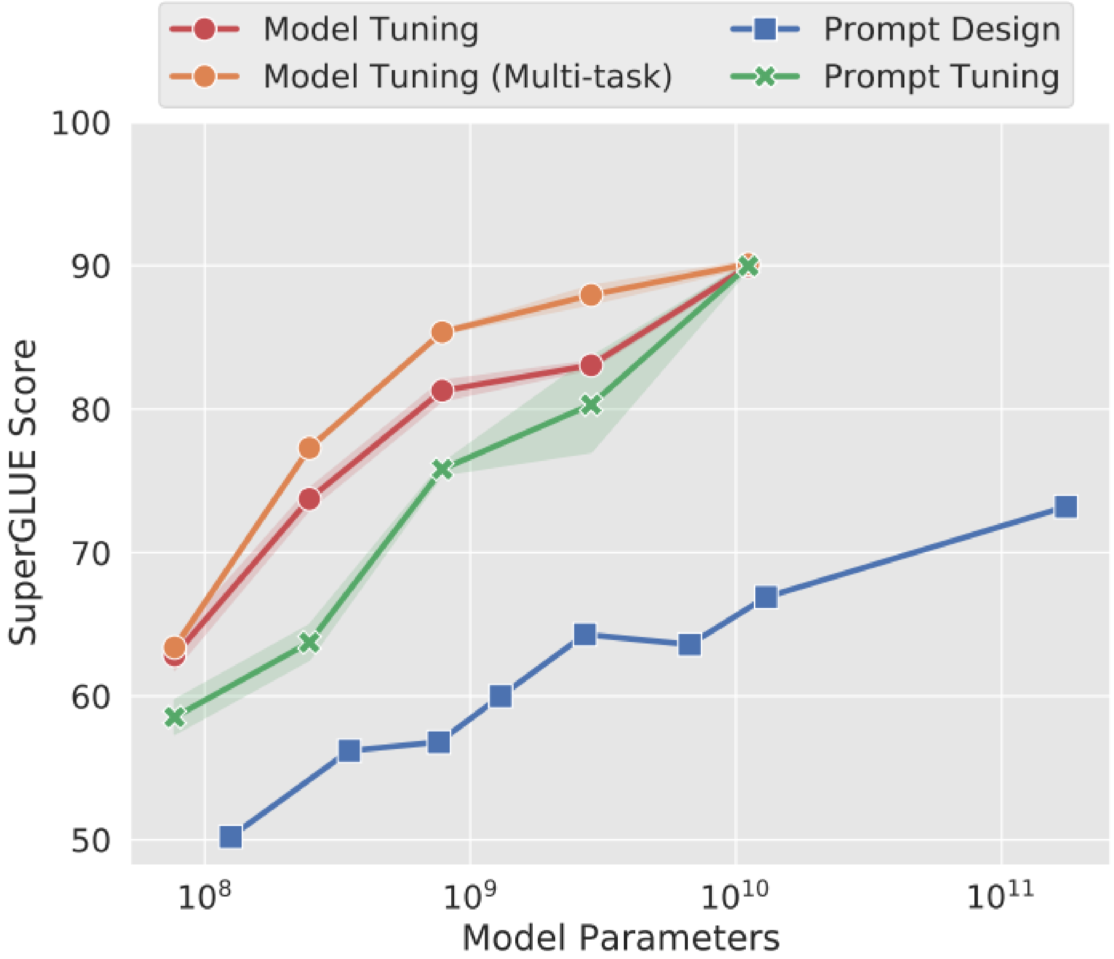


(a) Prompt length



(b) Prompt initialization

Результаты



Стандартная T5 - высокая производительность, но требует хранения отдельных копий модели для каждой конечной задачи.

Prompt tuned T5 повторно использует одну модель для всех задач.

Подход значительно превосходит несколько вариантов fewshot prompt для GPT-3.

Показано среднее значение и стандартное отклонение за 3 прогона для tuning'а

Результаты: ансамблирование

Dataset	Metric	Average	Best	Ensemble
BoolQ	acc.	91.1	91.3	**91.7**
CB	acc./F1	99.3 / 99.0	100.00 / 100.00	**100.0 / 100.0**
COPA	acc.	98.8	100.0	**100.0**
MultiRC	EM/$F1_a$	65.7 / 88.7	66.3 / 89.0	**67.1 / 89.4**
ReCoRD	EM/F1	92.7 / 93.4	92.9 / 93.5	**93.2 / 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**

Table 3: Performance of a five-prompt ensemble built from a single frozen T5-XXL model exceeds both the average and the best among the five prompts.

PPT подход

PT ~ full model tuning

!

PT << few-shot learning

PPT подход

PT ~ full model tuning   !   PT << few-shot learning

Решение: предобучать soft prompts,
добавив их в стадию предобучения

PPT подход

PT ~ full model tuning   !   PT << few-shot learning

Решение: предобучать soft prompts,
добавив их в стадию предобучения

+ для обощаемости похожие задачи классификации сводить в единую и предобучать под неё

PPT подход

(a) Full-Data

(b) Few-Shot

Figure 2: Comparison between PT and FT. The tuned prompt is composed of 100 learnable embeddings whose dimensions are the same as the token embeddings of PLMs (4096 dimensions). All these results are based on 11B PLMs T5 and CPM-2. FT needs to optimize all 11B parameters, while PT only trains about 410K prompt parameters.

PPT подход

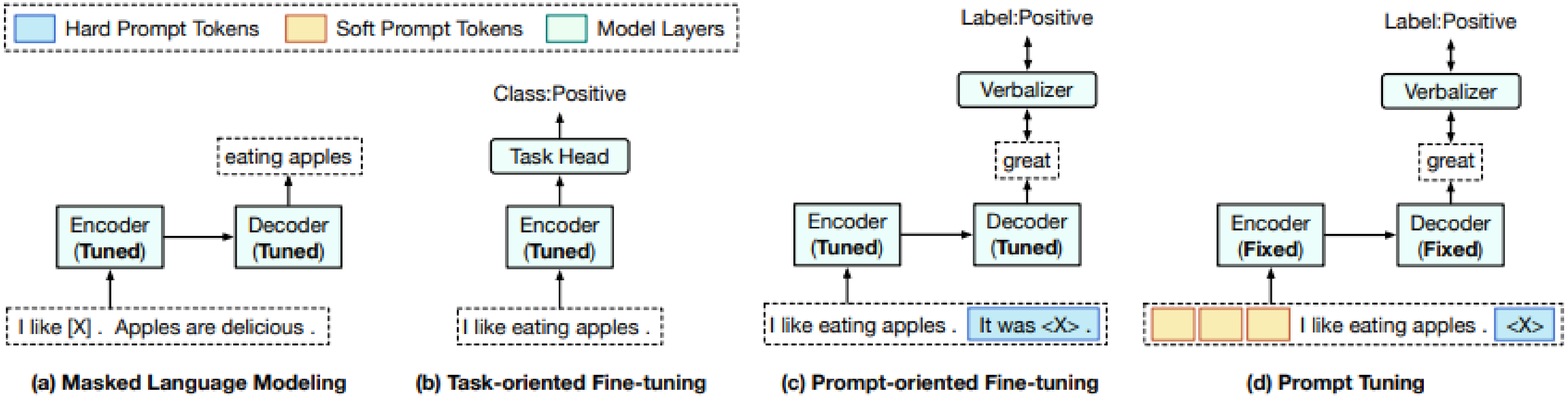


Figure 1: Paradigms of pre-training (masked language modeling), full-model tuning (task-oriented fine-tuning and prompt-oriented fine-tuning), and prompt tuning. The verbalizer is a function to map task labels to concrete words. ⟨X⟩ means the mask of typical pre-trained encoder-decoder models

PPT подход

$\{T_1, T_2, \dots, T_m\}$ – набор задач

T_i = $\{PVP_i^1, PVP_i^2, \dots, PVP_i^{n_i}\}$

PPT подход

$\{T_1, T_2, \ldots, T_m\}$ – набор задач

T_i = $\{PVP_i^1, PVP_i^2, \ldots, PVP_i^{n_i}\}$

$PVP_i^k = (f_i^k, v_i^k)$

f – mapping pattern, v - vocab

PPT подход

$\{T_1, T_2, \dots, T_m\}$ – набор задач

T_i = $\{PVP_i^1, PVP_i^2, \dots, PVP_i^{n_i}\}$

$$PVP_i^k = (f_i^k, v_i^k)$$

f – mapping pattern, v - vocab

$$PVP_i^{pre} = (f_i^{pre}, v_i^{pre})$$

PPT подход

$\{T_1, T_2, \dots, T_m\}$ – набор задач

T_i = $\{PVP_i^1, PVP_i^2, \dots, PVP_i^{n_i}\}$

$PVP_i^k = (f_i^k, v_i^k)$

f – mapping pattern, v - vocab

$PVP_i^{pre} = (f_i^{pre}, v_i^{pre})$

$\{P_1, P_2, \dots, P_m\}$ - prompts

PPT подход: setup для экспериментов

Chinese and English

PT on CPM-2 vs T5

PT on T5-XXL vs T5

PPT подход: setup для экспериментов

Chinese and English

PT on CPM-2 vs T5

PT on T5-XXL vs T5

- 100 soft tokens
- 410K params in PT (vs 11B FT)

PPT подход: итоги

			English Tasks						
	Model	Method	SST-2 Acc.	SST-5 Acc.	RACE-m Acc.	RACE-h Acc.	BoolQ Acc.	RTE Acc.	CB F1
FT (11B)	T5-Small	-	$72.8_{3.1}$	$31.1_{0.4}$	$26.4_{0.6}$	$26.3_{0.5}$	$59.2_{0.6}$	$54.0_{1.7}$	$70.1_{4.6}$
	T5-Base	-	$74.6_{2.7}$	$28.8_{1.8}$	$27.2_{0.5}$	$26.7_{0.2}$	$61.9_{2.1}$	$56.1_{2.3}$	$70.4_{2.6}$
	T5-Large	-	$89.1_{2.2}$	$42.4_{1.2}$	$48.2_{1.6}$	$43.2_{1.7}$	$74.6_{0.9}$	$64.4_{3.4}$	$82.3_{2.2}$
	T5-XL	-	$89.6_{3.2}$	$38.4_{5.1}$	$55.0_{2.8}$	$50.9_{2.6}$	$77.2_{2.1}$	$62.3_{6.8}$	$81.9_{9.0}$
	T5-XXL	-	$91.4_{0.8}$	$40.6_{2.0}$	$\mathbf{62.9_{3.9}}$	$\mathbf{54.8_{3.0}}$	$80.8_{2.4}$	$64.1_{2.0}$	$\mathbf{86.5_{5.3}}$
PT (410K)	T5-XXL	Vanilla PT	$70.5_{15.5}$	$32.3_{8.3}$	$34.7_{8.2}$	$31.6_{3.5}$	$61.0_{5.3}$	$53.5_{3.5}$	$50.7_{4.1}$
		Hybrid PT	$87.6_{6.6}$	$40.9_{2.7}$	$53.5_{8.2}$	$44.2_{6.4}$	$79.8_{1.5}$	$56.8_{2.6}$	$66.5_{7.2}$
		LM Adaption	$77.6_{7.5}$	$36.2_{3.6}$	$27.3_{0.2}$	$26.5_{0.4}$	$62.0_{0.3}$	$55.3_{1.0}$	$61.2_{1.7}$
		PPT	$93.5_{0.3}$	$\underline{\mathbf{50.2_{0.7}}}$	$60.0_{1.2}$	$\underline{53.0_{0.4}}$	$66.43_{5.7}$	$58.9_{1.6}$	$71.2_{6.2}$
		Hybrid PPT	$93.8_{0.1}$	$50.1_{0.5}$	$\underline{62.5_{0.9}}$	$52.2_{0.7}$	$\underline{\mathbf{82.0_{1.0}}}$	$59.8_{3.2}$	$73.2_{7.0}$
		Unified PPT	$\underline{\mathbf{94.4_{0.3}}}$	$46.0_{1.3}$	$58.0_{0.9}$	$49.9_{1.3}$	$76.0_{2.7}$	$\underline{\mathbf{65.8_{2.1}}}$	$\underline{82.2_{5.4}}$

			Chinese Tasks						
	Model	Method	ChnSent Acc.	Amazon Acc.	CCPM Acc.	C^3 Acc.	LCQMC Acc.	CMNLI Acc.	OCNLI Acc.
FT (11B)	mT5-Small	-	$76.1_{2.6}$	$29.9_{1.9}$	$31.9_{1.2}$	$29.6_{0.5}$	$52.4_{2.5}$	$36.5_{0.2}$	$34.9_{1.3}$
	mT5-Base	-	$78.2_{0.6}$	$36.4_{0.9}$	$40.4_{6.8}$	$29.4_{0.6}$	$50.9_{1.0}$	$36.3_{0.5}$	$35.4_{0.6}$
	mT5-Large	-	$79.1_{0.6}$	$31.0_{1.4}$	$46.0_{4.0}$	$29.9_{0.8}$	$52.1_{0.6}$	$35.8_{1.2}$	$35.2_{1.1}$
	mT5-XL	-	$82.7_{2.6}$	$35.5_{1.7}$	$68.3_{5.1}$	$29.7_{1.2}$	$52.9_{2.4}$	$36.8_{1.6}$	$35.6_{0.5}$
	mT5-XXL	-	$83.6_{1.5}$	$42.1_{0.8}$	$79.7_{1.1}$	$37.2_{3.3}$	$53.1_{1.0}$	$39.0_{0.4}$	$37.4_{1.2}$
	CPM-2	-	$86.1_{1.8}$	$42.5_{2.0}$	$81.8_{1.6}$	$38.4_{3.7}$	$58.8_{1.8}$	$40.7_{1.0}$	$38.5_{1.5}$
PT (410K)	CPM-2	Vanilla PT	$62.1_{3.1}$	$30.3_{4.8}$	$31.0_{9.7}$	$28.2_{0.4}$	$51.5_{3.4}$	$35.4_{0.5}$	$37.0_{0.5}$
		Hybrid PT	$79.2_{4.0}$	$39.1_{3.8}$	$46.6_{15.0}$	$29.2_{0.5}$	$54.6_{2.3}$	$37.1_{0.6}$	$37.8_{1.4}$
		LM Adaption	$74.3_{5.2}$	$35.2_{2.4}$	$33.7_{12.8}$	$30.2_{1.5}$	$51.4_{2.9}$	$35.1_{0.3}$	$38.0_{1.1}$
		PPT	$90.1_{0.8}$	$48.6_{0.6}$	$\underline{\mathbf{85.4_{0.6}}}$	$43.8_{2.2}$	$59.1_{0.6}$	$\underline{\mathbf{43.0_{0.5}}}$	$40.1_{0.4}$
		Hybrid PPT	$89.5_{0.3}$	$\underline{\mathbf{48.8_{2.0}}}$	$83.9_{0.5}$	$46.0_{0.5}$	$\underline{\mathbf{67.3_{0.9}}}$	$41.3_{0.8}$	$38.7_{0.6}$
		Unified PPT	$\underline{\mathbf{90.7_{0.2}}}$	$44.6_{1.1}$	$83.4_{0.9}$	$\underline{\mathbf{50.2_{0.6}}}$	$55.0_{0.4}$	$40.6_{0.4}$	$\underline{\mathbf{41.5_{1.5}}}$