

# LoRA

Голован Сергей  
Илюхов Алексей  
Аюпов Шамиль  
Шошин Борис

# Problem statement

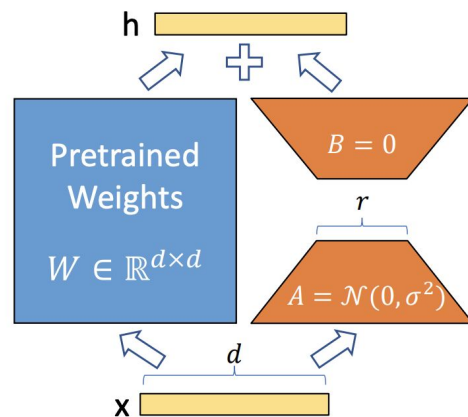
- Представим, что у нас есть модель  $P_{\Phi}(y|x)$  с параметрами  $\Phi$
- Датасет для обучения  $\mathcal{Z} = \{(x_i, y_i)\}_{i=1, \dots, N}$
- Обновление  $\Phi_0 + \Delta\Phi$
- 
- $$\max_{\Phi} \sum_{(x,y) \in \mathcal{Z}} \sum_{t=1}^{|y|} \log(P_{\Phi}(y_t|x, y_{<t}))$$
- 
- Что делается в статье
- $$\max_{\Theta} \sum_{(x,y) \in \mathcal{Z}} \sum_{t=1}^{|y|} \log(p_{\Phi_0 + \Delta\Phi(\Theta)}(y_t|x, y_{<t}))$$

## Метод

$$W_0 + \Delta W = W_0 + BA, \text{ где } W_0 \in \mathbb{R}^{d \times k}, B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$$

$$r \ll \min(d, k)$$

$$h = W_0 x + \Delta W x = W_0 x + BAx$$



# Эксперименты

Model & Method	# Trainable Parameters	E2E NLG Challenge				
		BLEU	NIST	MET	ROUGE-L	CIDEr
GPT-2 M (FT)*	354.92M	68.2	8.62	46.2	71.0	2.47
GPT-2 M (Adapter <sup>L</sup> )*	0.37M	66.3	8.41	45.0	69.8	2.40
GPT-2 M (Adapter <sup>L</sup> )*	11.09M	68.9	8.71	46.1	71.3	2.47
GPT-2 M (Adapter <sup>H</sup> )	11.09M	67.3 $\pm$ .6	8.50 $\pm$ .07	46.0 $\pm$ .2	70.7 $\pm$ .2	2.44 $\pm$ .01
GPT-2 M (FT <sup>Top2</sup> )*	25.19M	68.1	8.59	46.0	70.8	2.41
GPT-2 M (PreLayer)*	0.35M	69.7	8.81	46.1	71.4	2.49
GPT-2 M (LoRA)	0.35M	<b>70.4</b> $\pm$ .1	<b>8.85</b> $\pm$ .02	<b>46.8</b> $\pm$ .2	<b>71.8</b> $\pm$ .1	<b>2.53</b> $\pm$ .02
GPT-2 L (FT)*	774.03M	68.5	8.78	46.0	69.9	2.45
GPT-2 L (Adapter <sup>L</sup> )	0.88M	69.1 $\pm$ .1	8.68 $\pm$ .03	46.3 $\pm$ .0	71.4 $\pm$ .2	<b>2.49</b> $\pm$ .0
GPT-2 L (Adapter <sup>L</sup> )	23.00M	68.9 $\pm$ .3	8.70 $\pm$ .04	46.1 $\pm$ .1	71.3 $\pm$ .2	2.45 $\pm$ .02
GPT-2 L (PreLayer)*	0.77M	70.3	8.85	46.2	71.7	2.47
GPT-2 L (LoRA)	0.77M	<b>70.4</b> $\pm$ .1	<b>8.89</b> $\pm$ .02	<b>46.8</b> $\pm$ .2	<b>72.0</b> $\pm$ .2	2.47 $\pm$ .02

Table 3: GPT-2 medium (M) and large (L) with different adaptation methods on the E2E NLG Challenge. For all metrics, higher is better. LoRA outperforms several baselines with comparable or fewer trainable parameters. Confidence intervals are shown for experiments we ran. \* indicates numbers published in prior works.

Model&Method	# Trainable Parameters	WikiSQL	MNLI-m	SAMSum
		Acc. (%)	Acc. (%)	R1/R2/RL
GPT-3 (FT)	175,255.8M	<b>73.8</b>	89.5	52.0/28.0/44.5
GPT-3 (BitFit)	14.2M	71.3	91.0	51.3/27.4/43.5
GPT-3 (PreEmbed)	3.2M	63.1	88.6	48.3/24.2/40.5
GPT-3 (PreLayer)	20.2M	70.1	89.5	50.8/27.3/43.5
GPT-3 (Adapter <sup>H</sup> )	7.1M	71.9	89.8	53.0/28.9/44.8
GPT-3 (Adapter <sup>H</sup> )	40.1M	73.2	<b>91.5</b>	53.2/29.0/45.1
GPT-3 (LoRA)	4.7M	73.4	<b>91.7</b>	<b>53.8/29.8/45.9</b>
GPT-3 (LoRA)	37.7M	<b>74.0</b>	<b>91.6</b>	53.4/29.2/45.1

Table 4: Performance of different adaptation methods on GPT-3 175B. We report the logical form validation accuracy on WikiSQL, validation accuracy on MultiNLI-matched, and Rouge-1/2/L on SAMSum. LoRA performs better than prior approaches, including full fine-tuning. The results on WikiSQL have a fluctuation around  $\pm 0.5\%$ , MNLI-m around  $\pm 0.1\%$ , and SAMSum around  $\pm 0.2/\pm 0.2/\pm 0.1$  for the three metrics.

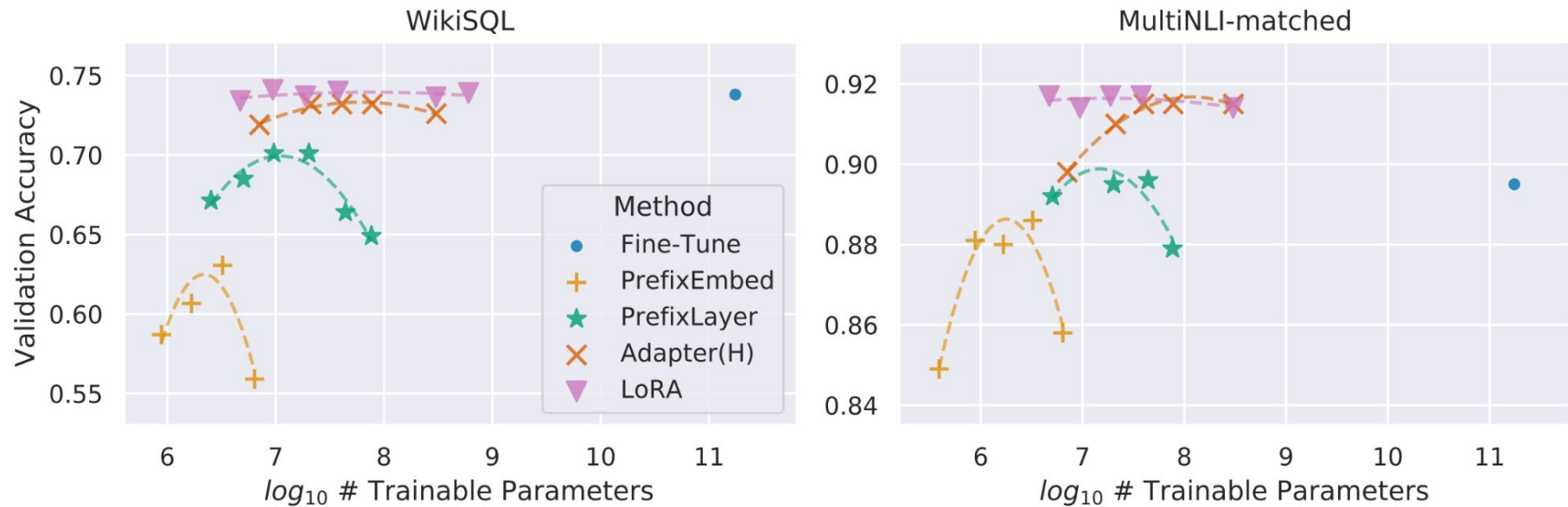


Figure 2: GPT-3 175B validation accuracy vs. number of trainable parameters of several adaptation methods on WikiSQL and MNLI-matched. LoRA exhibits better scalability and task performance. See Section F.2 for more details on the plotted data points.