

Parameter Efficient Fine-Tuning

Бузмаков Василий Олегович, БПМИ213

План

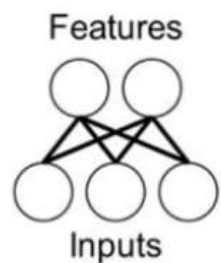
1. Что хотим?
2. Базовые методы дообучения модели
3. PEFT

Так называемая база

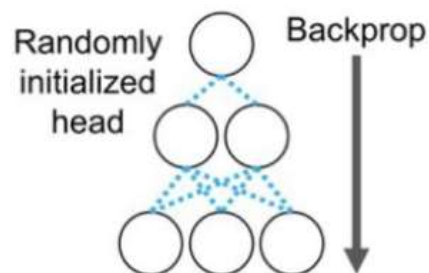
ID – in distribution

OOD – out of distribution

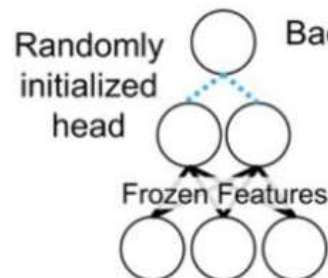
Pretraining



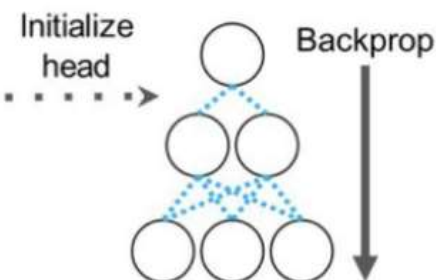
(a) Fine-tuning



(b) Linear probing



(c) LP-FT



ID test

85.1%

82.9%

85.7%

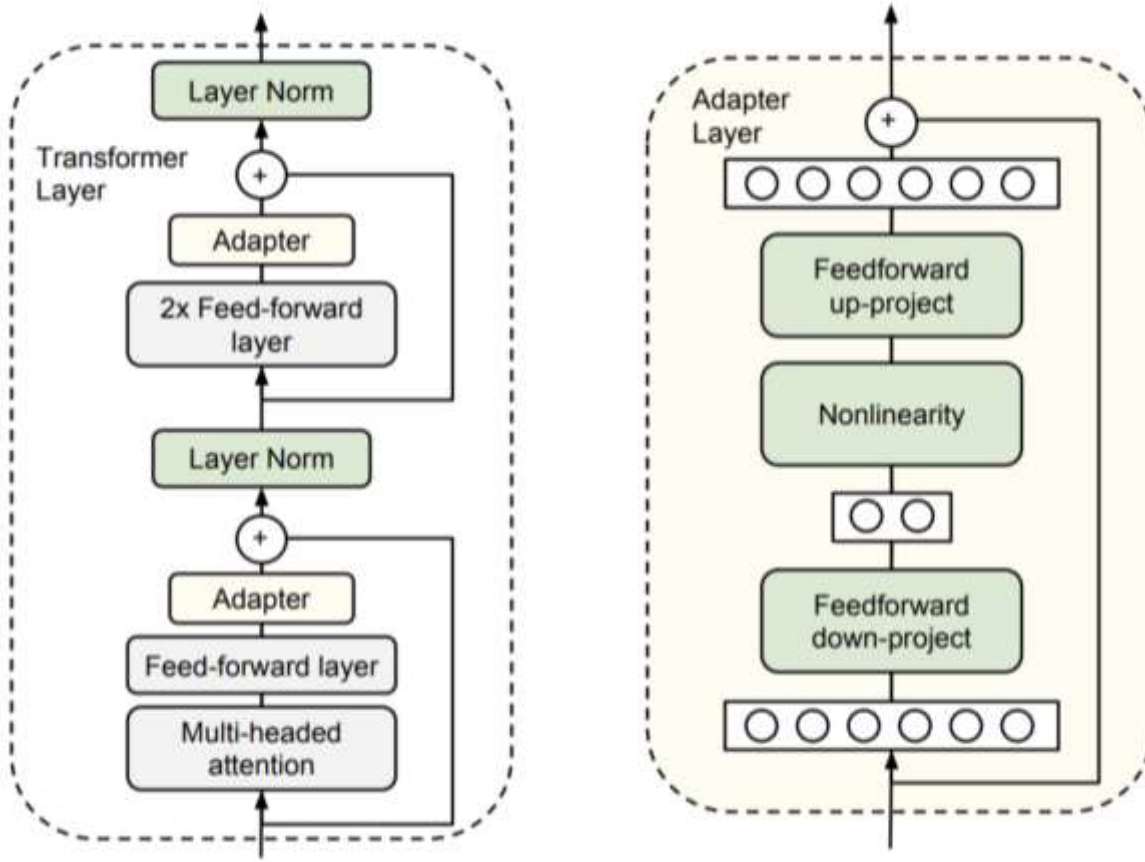
OOD test

59.3%

66.2%

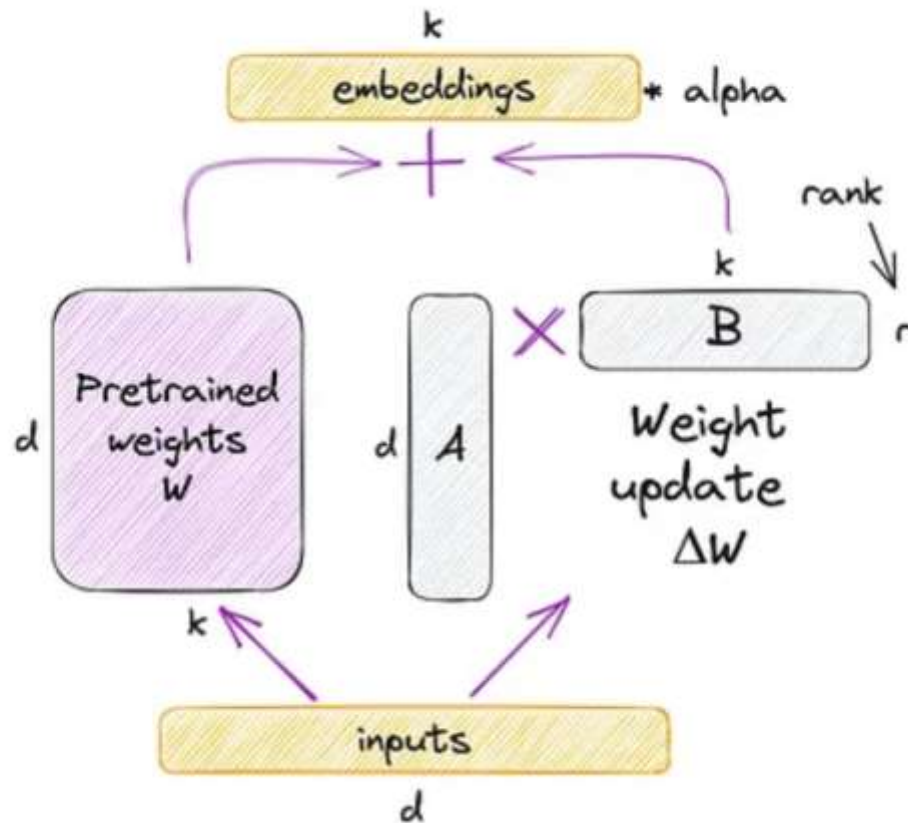
68.9%

Adapters



Добавляем Adapter слой в блоки трансформера. Обучаем только его. Можем подбирать размерность, в которую переводит feedforward-down.

Low Rank Adaptation (LoRa)



Можем подбирать параметр r – ранг матриц A и B . B изначально инициализуется 0.

Приближаем матрицы весов матрицами меньшего ранга

BitFit (Bias-Terms Fine Tuning)

Обновляем только bias в линейных слоях модели

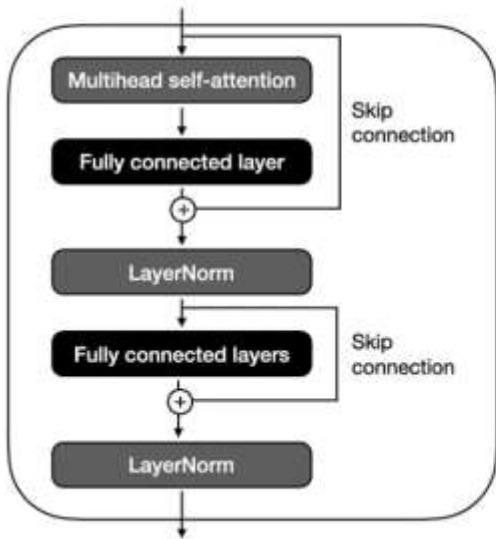
$$\mathbf{Q}^{m,\ell}(\mathbf{x}) = \mathbf{W}_q^{m,\ell} \mathbf{x} + \mathbf{b}_q^{m,\ell}$$

$$\mathbf{K}^{m,\ell}(\mathbf{x}) = \mathbf{W}_k^{m,\ell} \mathbf{x} + \mathbf{b}_k^{m,\ell}$$

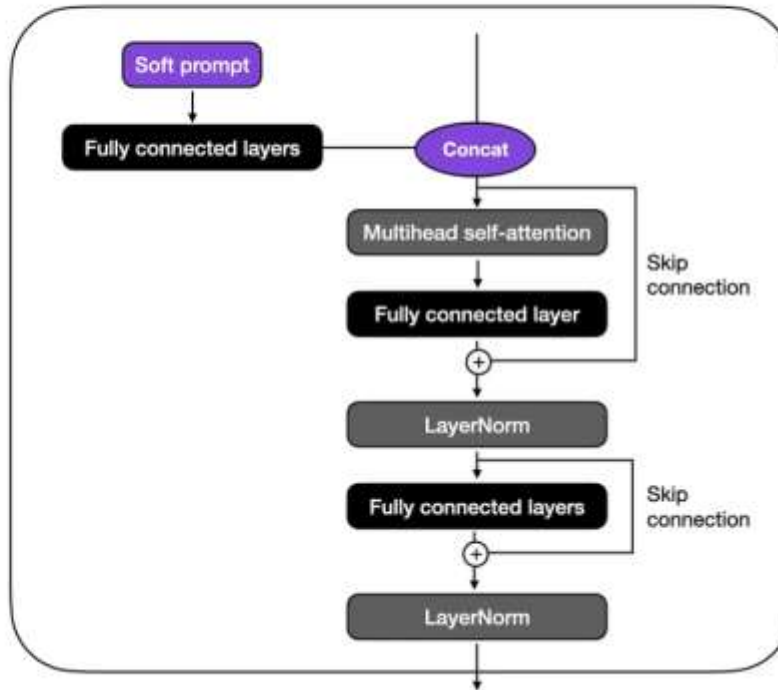
$$\mathbf{V}^{m,\ell}(\mathbf{x}) = \mathbf{W}_v^{m,\ell} \mathbf{x} + \mathbf{b}_v^{m,\ell}$$

Prefix tuning

REGULAR TRANSFORMER BLOCK



TRANSFORMER BLOCK WITH PREFIX



Soft prompt – обучаемый тензор, который конкатенируется с эмбедами текста при обучении.

Идея prefix tuning заключается в добавлении обучаемых тензоров к каждому блоку трансформера. Кроме того, мы прогоняем soft prompt через полносвязный слой.

Результаты

	50k examples					22k examples										82k examples					
	E2E					WebNLG										DART					
	BLEU	NIST	MET	R-L	CIDEr	BLEU			MET			TER ↓				BLEU	MET	TER ↓	Mover	BERT	BLEURT
						S	U	A	S	U	A	S	U	A							
GPT-2 _{MEDIUM}																					
FINE-TUNE	68.2	8.62	46.2	71.0	2.47	64.2	27.7	46.5	0.45	0.30	0.38	0.33	0.76	0.53		46.2	0.39	0.46	0.50	0.94	0.39
FT-TOP2	68.1	8.59	46.0	70.8	2.41	53.6	18.9	36.0	0.38	0.23	0.31	0.49	0.99	0.72		41.0	0.34	0.56	0.43	0.93	0.21
ADAPTER(3%)	68.9	8.71	46.1	71.3	2.47	60.4	48.3	54.9	0.43	0.38	0.41	0.35	0.45	0.39		45.2	0.38	0.46	0.50	0.94	0.39
ADAPTER(0.1%)	66.3	8.41	45.0	69.8	2.40	54.5	45.1	50.2	0.39	0.36	0.38	0.40	0.46	0.43		42.4	0.36	0.48	0.47	0.94	0.33
PREFIX(0.1%)	69.7	8.81	46.1	71.4	2.49	62.9	45.6	55.1	0.44	0.38	0.41	0.35	0.49	0.41		46.4	0.38	0.46	0.50	0.94	0.39
GPT-2 _{LARGE}																					
FINE-TUNE	68.5	8.78	46.0	69.9	2.45	65.3	43.1	55.5	0.46	0.38	0.42	0.33	0.53	0.42		47.0	0.39	0.46	0.51	0.94	0.40
Prefix	70.3	8.85	46.2	71.7	2.47	63.4	47.7	56.3	0.45	0.39	0.42	0.34	0.48	0.40		46.7	0.39	0.45	0.51	0.94	0.40

Результаты

Model & Method	# Trainable Parameters	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	Avg.
RoB _{base} (FT)*	125.0M	87.6	94.8	90.2	63.6	92.8	91.9	78.7	91.2	86.4
RoB _{base} (BitFit)*	0.1M	84.7	93.7	92.7	62.0	91.8	84.0	81.5	90.8	85.2
RoB _{base} (Adpt ^D)*	0.3M	87.1 \pm 0	94.2 \pm 1	88.5 \pm 1.1	60.8 \pm 4	93.1 \pm 1	90.2 \pm 0	71.5 \pm 2.7	89.7 \pm 3	84.4
RoB _{base} (Adpt ^D)*	0.9M	87.3 \pm 1	94.7 \pm 3	88.4 \pm 1	62.6 \pm 9	93.0 \pm 2	90.6 \pm 0	75.9 \pm 2.2	90.3 \pm 1	85.4
RoB _{base} (LoRA)	0.3M	87.5 \pm 3	95.1\pm2	89.7 \pm 7	63.4 \pm 1.2	93.3\pm3	90.8 \pm 1	86.6\pm7	91.5\pm2	87.2
RoB _{large} (FT)*	355.0M	90.2	96.4	90.9	68.0	94.7	92.2	86.6	92.4	88.9
RoB _{large} (LoRA)	0.8M	90.6\pm2	96.2 \pm 5	90.9\pm1.2	68.2\pm1.9	94.9\pm3	91.6 \pm 1	87.4\pm2.5	92.6\pm2	89.0
RoB _{large} (Adpt ^P)†	3.0M	90.2 \pm 3	96.1 \pm 3	90.2 \pm 7	68.3\pm1.0	94.8\pm2	91.9\pm1	83.8 \pm 2.9	92.1 \pm 7	88.4
RoB _{large} (Adpt ^P)†	0.8M	90.5\pm3	96.6\pm2	89.7 \pm 1.2	67.8 \pm 2.5	94.8\pm3	91.7 \pm 2	80.1 \pm 2.9	91.9 \pm 4	87.9
RoB _{large} (Adpt ^H)†	6.0M	89.9 \pm 5	96.2 \pm 3	88.7 \pm 2.9	66.5 \pm 4.4	94.7 \pm 2	92.1 \pm 1	83.4 \pm 1.1	91.0 \pm 1.7	87.8
RoB _{large} (Adpt ^H)†	0.8M	90.3 \pm 3	96.3 \pm 5	87.7 \pm 1.7	66.3 \pm 2.0	94.7 \pm 2	91.5 \pm 1	72.9 \pm 2.9	91.5 \pm 5	86.4
RoB _{large} (LoRA)†	0.8M	90.6\pm2	96.2 \pm 5	90.2\pm1.0	68.2 \pm 1.9	94.8\pm3	91.6 \pm 2	85.2\pm1.1	92.3\pm5	88.6
DeB _{XXL} (FT)*	1500.0M	91.8	97.2	92.0	72.0	96.0	92.7	93.9	92.9	91.1
DeB _{XXL} (LoRA)	4.7M	91.9\pm2	96.9 \pm 2	92.6\pm6	72.4\pm1.1	96.0\pm1	92.9\pm1	94.9\pm4	93.0\pm2	91.3

- <https://arxiv.org/pdf/2303.15647.pdf>
- https://www.youtube.com/watch?v=Ylj9M_Ufy64&list=PLEwK9wdS5g0pc4NeOQqGLPcxmBHGUjnWB&index=16
- <https://arxiv.org/pdf/2106.09685.pdf>
- <https://magazine.sebastianraschka.com/p/understanding-parameter-efficient>
- <https://habr.com/ru/articles/791966/>