Skoltech

Parameter-Efficient Fine-Tuning (PEFT)

https://github.com/danyaaivanov/Team_Full_House_PEFT

Full house
Ivanov Danil
Shepelin Oleg
Gureeva Irena
Pyatkin Stanislav
Novikova Emiliya

Problem statement

An important approach in NLP is large-scale pretraining on general domain data and adaptation to particular tasks or domains

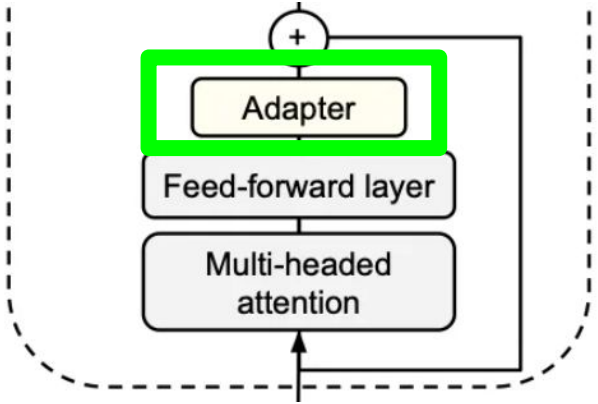
But full fine-tuning large-scale language models is often **prohibitively costly**

Model	Category	Size
GPT3 [55]	Causal decoder	175B
PanGU- α [75]	Causal decoder	207B
OPT [81]	Causal decoder	175B
PaLM [56]	Causal decoder	540B
BLOOM [69]	Causal decoder	176B
MT-NLG [97]	Causal decoder	530B
Gopher [59]	Causal decoder	280B
Chinchilla [34]	Causal decoder	70B
Galactica [35]	Causal decoder	120B
LaMDA [63]	Causal decoder	137B
Jurassic-1 [91]	Causal decoder	178B
LLaMA [57]	Causal decoder	65B
GLM-130B [83]	Prefix decoder	130B
T5 [73]	Encoder-decoder	11B

Two ways to solve it

Add **small neural modules** to PLMs and fine-tune only these modules for each task

Examples: adapter tuning, prefix tuning, promt tuning



Model the **incremental update** of the pre-trained weights in a parameter-efficient way

Examples: LoRA, AdaLoRA, diff pruning

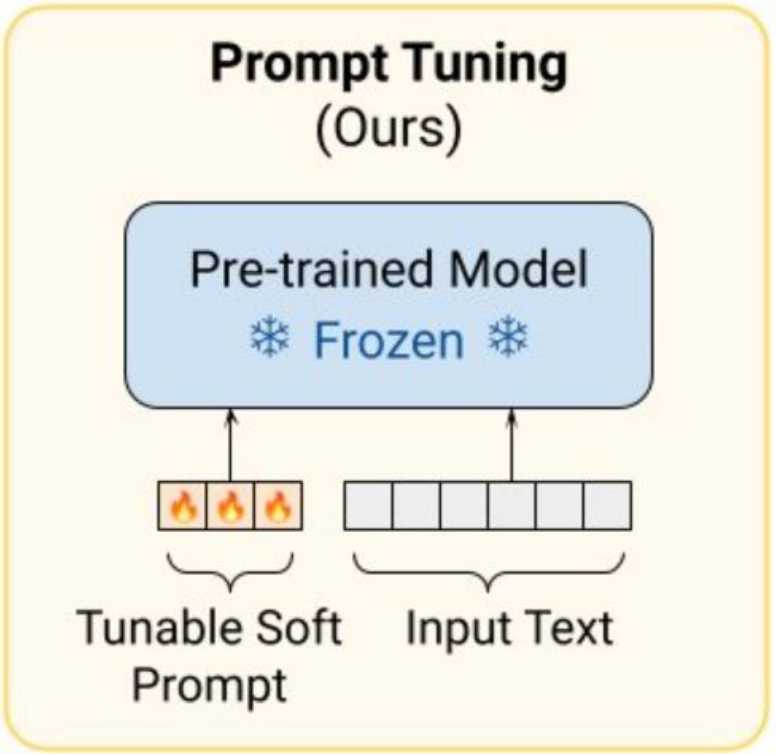
$$h = W_0 x + \Delta W x$$

Prompt/prefix tuning

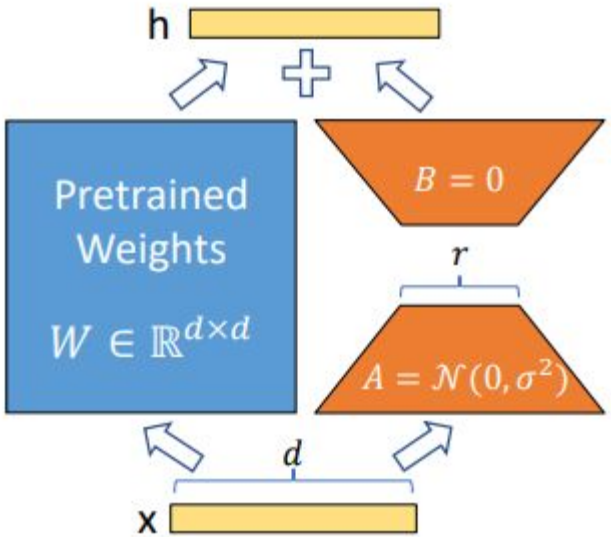
Only a **sequence of continuous task-specific vectors** is attached to the beginning of the input

Hard prompt tuning: change the **discrete** input tokens

Soft prompt tuning: add a **trainable tensor** to the input and/or to each transformer block (*prefix* tuning)



LoRA



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

$$B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$$

LoRA still has limitations as it prespecifies the rank **r of each incremental matrix** Δ **identical**

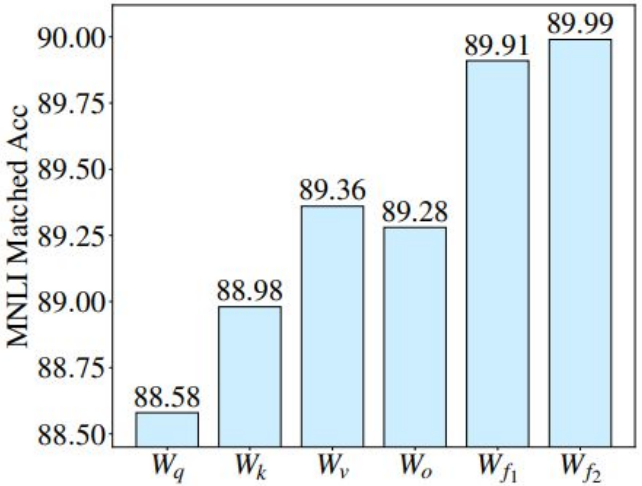
This ignores the fact that the **importance** of weight matrices varies significantly **across modules and layers**

How can we allocate the parameter budget **adaptively**?

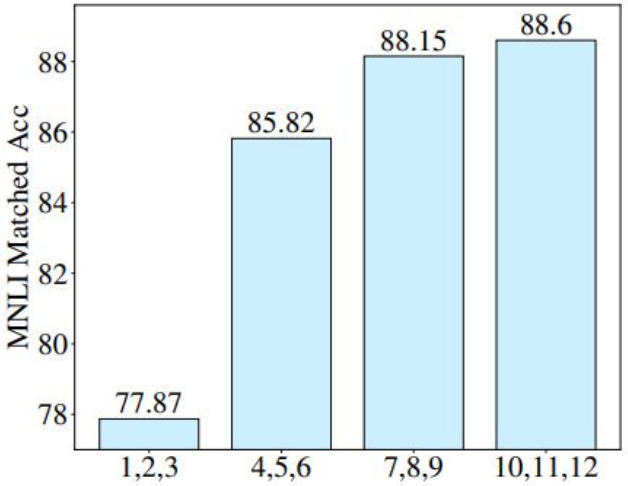
AdaLoRA

LoRA + Importance rank

$$W = W^{(0)} + \Delta = W^{(0)} + P\Lambda Q,$$



(a) Selected weight matrix



(b) Selected layers

Skoltech

QLoRA

Reaches 99.3% of the performance level of ChatGPT while only requiring 24 hours of finetuning on a single GPU

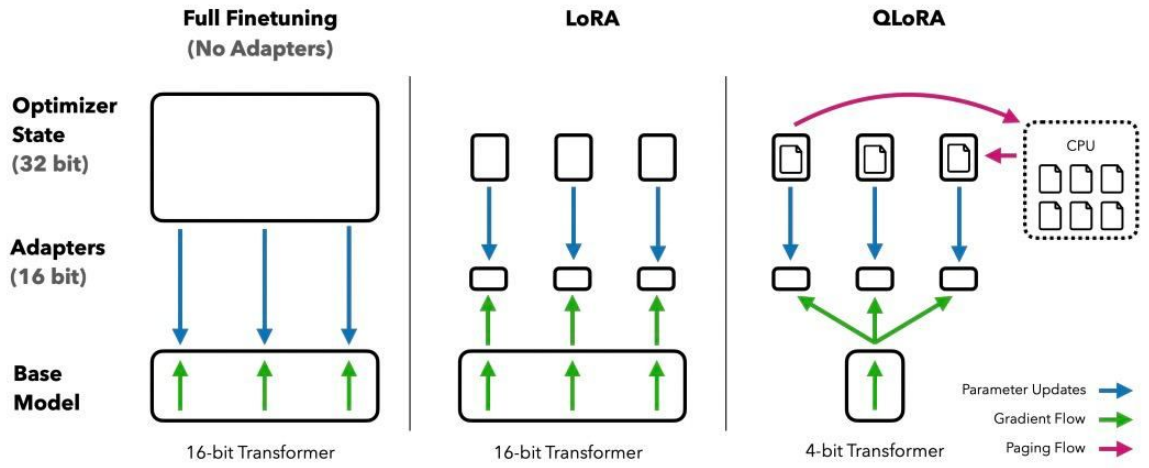


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

PEFT library

PEFT

State-of-the-art Parameter-Efficient Fine-Tuning (PEFT) methods

Supported methods:

1. LoRA: LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS
2. Prefix Tuning: Prefix-Tuning: Optimizing Continuous Prompts for Generation, P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks
3. P-Tuning: GPT Understands, Too
4. Prompt Tuning: The Power of Scale for Parameter-Efficient Prompt Tuning
5. AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning

Skoltech

Obtained results: Tables

The goal was to **summarize the content** of the tables

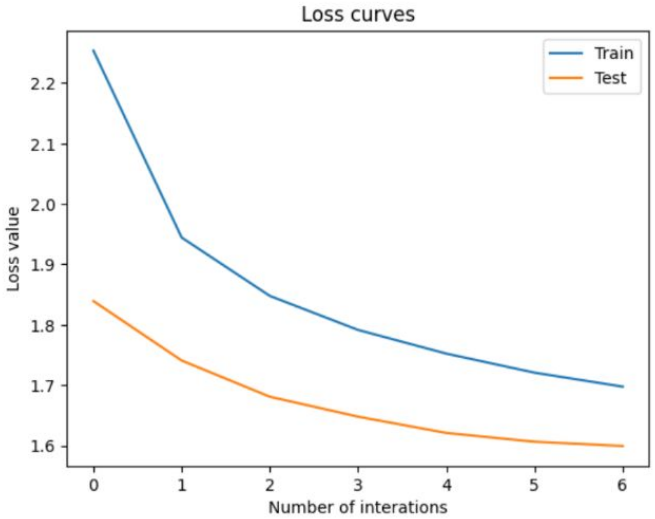
	table	summary
0	summarize: 0 Country: Russia, Medical device e...	Medical device expenditure per capita in the l...
1	summarize: 0 Country: Sweden, Number of cases:...	Cumulative number of coronavirus (COVID-19) ...
2	summarize: 0 radio company: Pandora Corporate,...	Leading online radio companies in the United S...
3	summarize: 0 Response: March 30, Share of resp...	Share of Indonesian population who would suppo...
4	summarize: 0 Year: 15 to 19 years, Number of c...	Number of births by age of mother in the Unite...
...	...	...
20251	summarize: 0 Year: 2017, Number of enterprises...	Number of enterprises in the manufacture of co...
20252	summarize: 0 Quarter: Q3 2019, Viewers in thou...	Quarterly reach of Dave television channel in ...
20253	summarize: 0 Year: 2019, Average ticket price ...	Average ticket price for New England Patriots ...
20254	summarize: 0 Year: 2025*, National debt to GDP...	Canada : National debt from 2015 to 2025 in re...
20255	summarize: 0 Country: China, Reserves in metri...	Reserves of tungsten worldwide in 2019 , by co...

20256 rows × 2 columns

Skoltech

Obtained results: Tables

LoRA on T5 model



	rouge-1	rouge-2	rouge-l
r	0.63	0.44	0.62
p	0.64	0.35	0.63
f	0.63	0.39	0.62

`bleu_score=0.03 on the evaluation dataset`

Example of the model output: 'Bulgaria : Ratio of government expenditure to gross domestic product (GDP) from 2015 to 2025 Ghana Ghana Ghana Ghana Ghana Ghana Bulgaria ... Bulgaria Rat Rat Rat India India India India India India India India India India India'

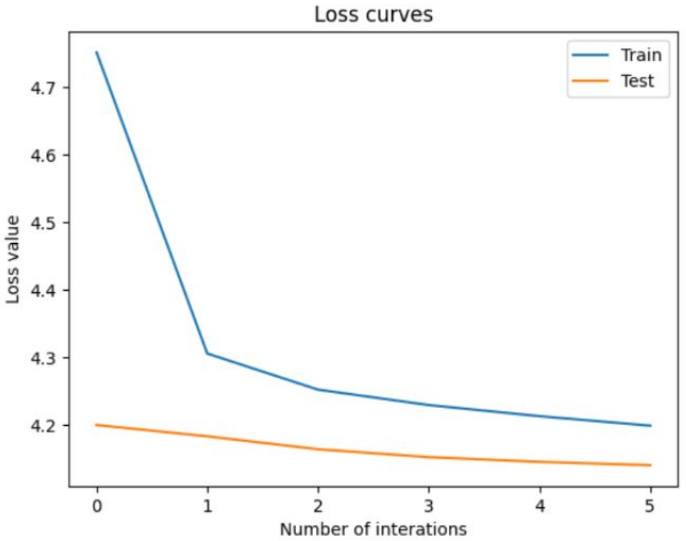
'...' denotes 'Bulgaria' repeated a lot of times.

Corresponding target: 'France : Ratio of government expenditure to gross domestic product (GDP) from 2015 to 2025'

Skoltech

Obtained results: Tables

AdaLoRA on BART model



	rouge-1	rouge-2	rouge-1
r	0.01	0.0	0.01
p	0.02	0.0	0.02
f	0.01	0.0	0.01

`bleu_score=0.26 on the evaluation dataset`

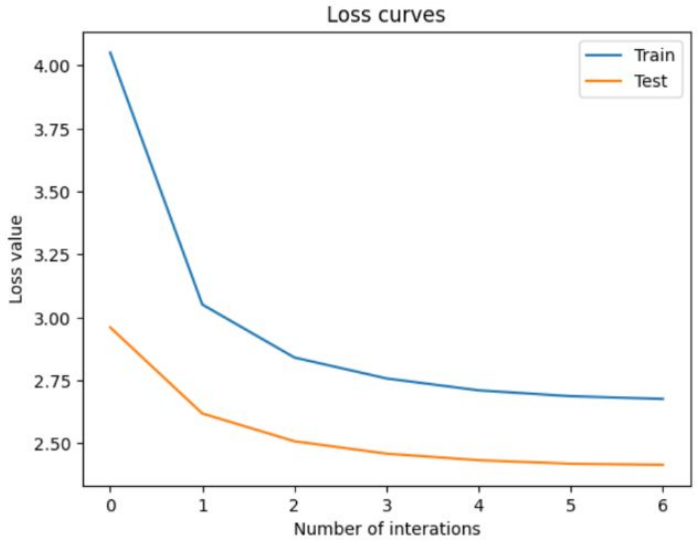
Example of the model output: 'both annuala well Empires iPad ganzberen special have hisatings lastX annualenen well well well well well well wellssssssssssssssssssssssssssssssssssssssssssssXsssssssssssssssssssssssssssssss semi semi semis semi much much much much much much much,,,,,,,,,,,'

Corresponding target: 'France : Ratio of government expenditure to gross domestic product (GDP) from 2015 to 2025'

Skoltech

Obtained results: Tables

Prefix Tuning on T5 model



	rouge-1	rouge-2	rouge-l
r	0.50	0.28	0.48
p	0.54	0.22	0.52
f	0.52	0.24	0.50

`bleu_score=0.04 on the evaluation dataset`

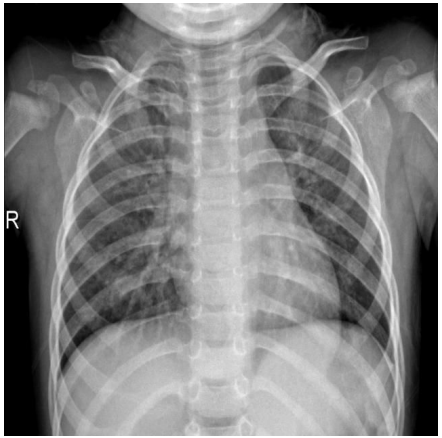
Example of the model output: 'Rat : Ratio of government expenditure to GDP domestic product (GDP) from 2015 to 2025 (Rat ... Rat'

'...' denotes 'Rat' repeated a lot of times.

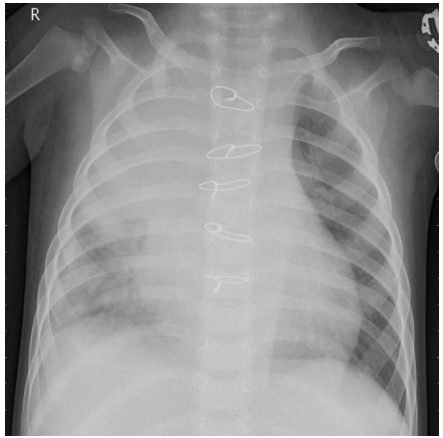
Corresponding target: 'France : Ratio of government expenditure to gross domestic product (GDP) from 2015 to 2025'

Skoltech

Obtained results: Image classification



Normal



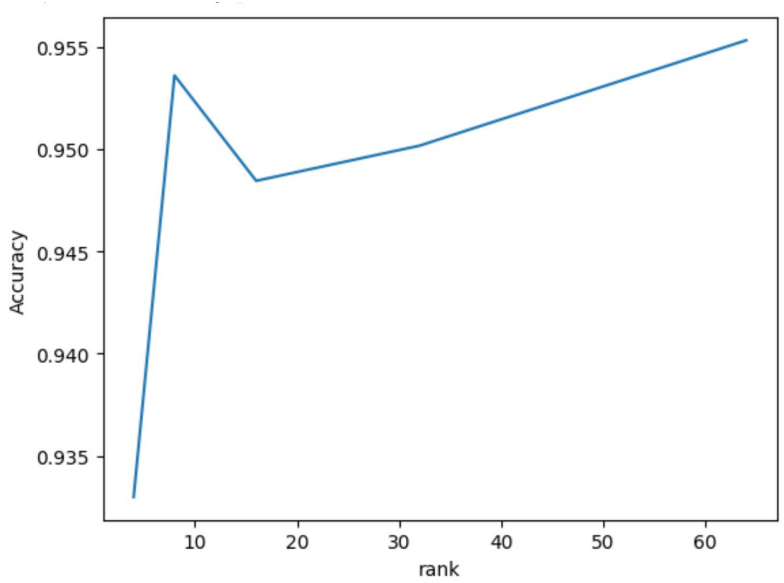
Pneumonia

Vision Transformer (ViT) model pre-trained on ImageNet-21k

Skoltech

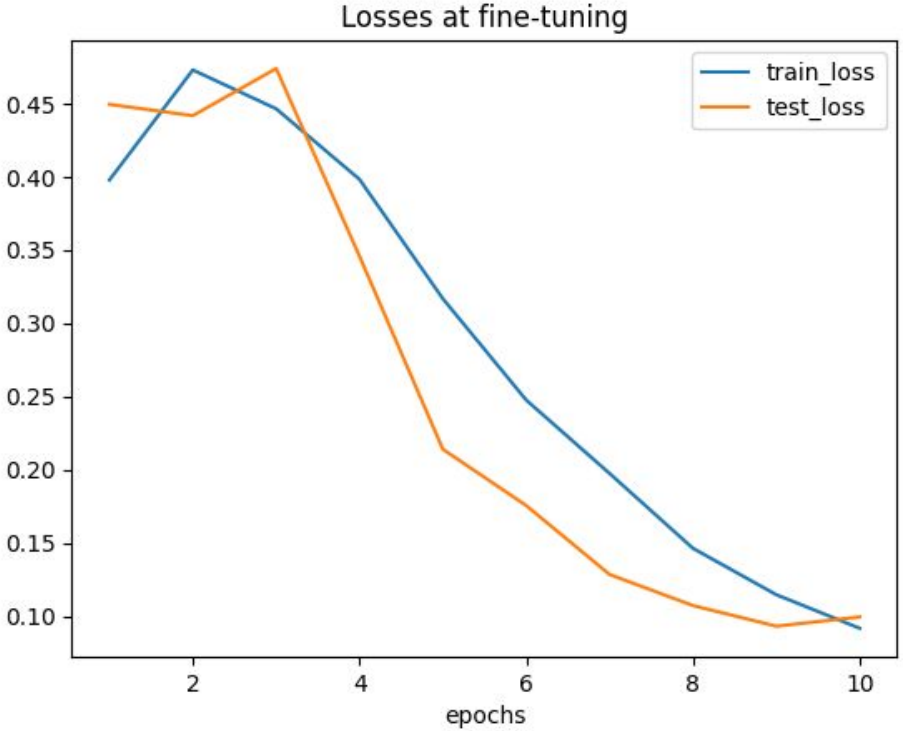
Obtained results: Image classification

Modules	Accuracy	Loss
query	0.917	0.307
value	0.929	0.157
key	0.931	0.171
query, value	0.939	0.166
query, key	**0.953**	0.135
value, key	0.932	0.151



Skoltech

Obtained results: Seq2Seq Generation



Hugging Face BigScience mt0-large model

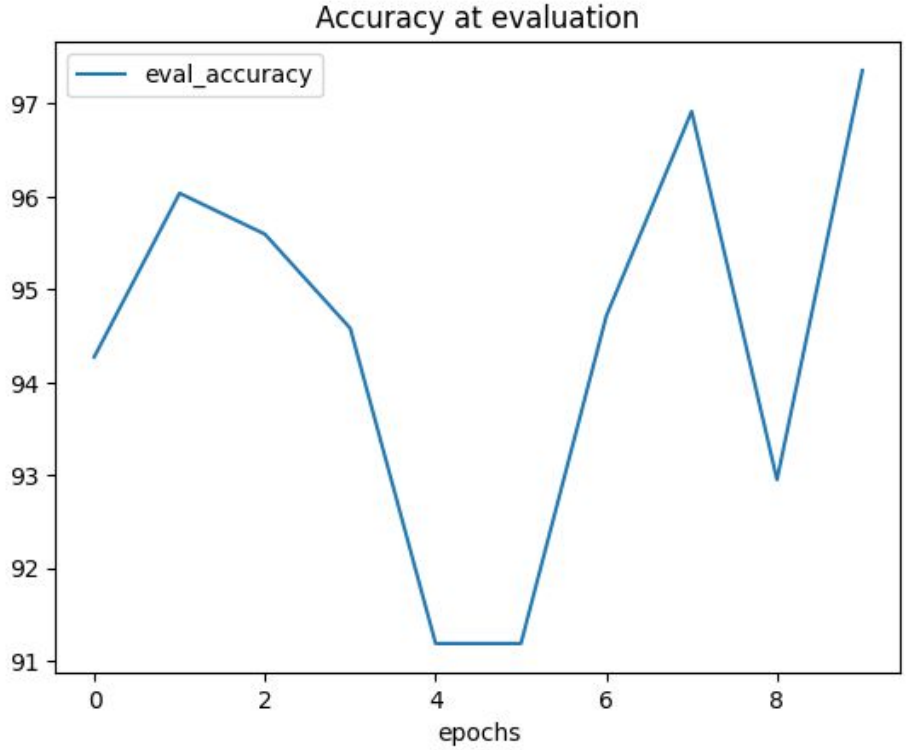
Total of 1.2B parameters. Total tuned parameters 0.2% (2.36M).

Pretrained originally on xP3 dataset

Fine-tuned on Financial Phrase Bank dataset.

Skoltech

Obtained results: Seq2Seq Generation



With just two epochs accuracy reached 90%+ levels and continued to climb higher tending to 10-th epoch.

Total fine-tuning and inference time was about 15 minutes.

Conclusion

- LoRA and Prefix Tuning on T5 model have proven themselves to give promising results on the table description task possibly leading to great performance, while AdaLoRA method applied to BART model requires further investigation
- For image classification, LoRA manages to achieve good accuracy with around 2.5% of parameters, getting 90%+ accuracy on new dataset from the second epoch.
- LoRA achieved peak accuracy from just very few epochs of fine-tuning while manipulating only a small fracture of less than 0.5% of parameters.