# Final Project in Numerical Linear Algebra

## Speeding up LoRA-FA with reasonable initialization and regularization

Maksim Komiakov  Sergey Karpukhin  Pavel Tikhomirov
Yulia Sergeeva  Pavel Bartenev

Skolkovo Institute of Science and Technology

2023

Skoltech

# Problem statement: Challenge of fine-tuning LLMs

We have some pretrained large language model defined by it's parameters $\theta_{old}$ and want to finetune (additionally train) it for downstream task.

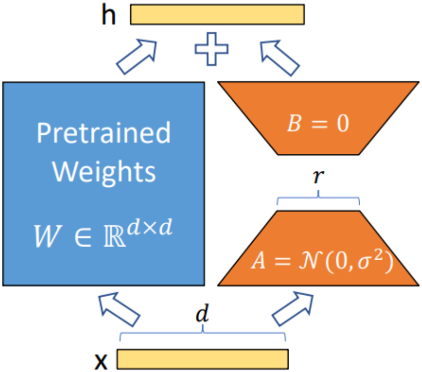
Possible solutions:

1. $\theta_{old} \to \theta_{new}$ - train all parameters - need a lot of compute!
2. $\theta_{new} = [\theta_{old} | \Delta\theta]$ - plug in trainable modules - affects latency!

Both approaches have their downsides.

Skoltech

# Low Rank Adaptation (2021)




$$W_0 + \Delta W = W_0 + BA$$

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

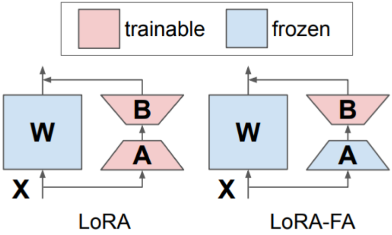
Instead of training full matrices of parameter $W$, updates can be constrained via **low-rank decomposition** of $\Delta W = BA$.

After training $\Delta W$ is simply merged with original weights **without introducing computational overhead** (and such operation can be done on the fly).

Skoltech

# LoRA-FA (2023)

One of many modifications of original LoRA.




**Main idea**: In low-rank approximation $AB$ **freeze** $A$ and train only projection to high-dim space $B$.
**Technical detail**: Randomly initialized $A$ is orthogonalized via $QR$-decomposition, only Q is frozen - some random orthogonal basis.
**Main advantage**: reduce number of trainable parameters - less memory, faster training, $\sim$ same quality

Skoltech

# Learning low-rank approximation

Both LoRA and LoRA-FA rely on somewhat random initialization of low-rank AB. But can we do better?

We try by using information from best-possible r-rank approximation of original weights $W$ - **Singular Value Decomposition**. Our intuition is that choosing this starting point might lead to faster convergence.

Skoltech

# Introduced modification

Initialization of LoFA-FA layer with SVD of original weight matrix $W_0$.

Algorithm is following:

1. compute $W_0 = U_r \Sigma_r V_r^T$
2. $A \leftarrow (U_r \Sigma_r)^T$
3. $B \leftarrow V_r$
4. freeze $A$

After that training procedure is very much same as for standard LoRA.

Skoltech

# Experiments

We experiment with fine-tuning of RoBERTa-base model (125 M. parameters) on various datasets (MRPC, COLA, RTE, STS-B) and fine-tune model for downstream classification task.

We compare our SVD initialization against original LoRA-FA. Our goal is to observe whether this initialization does any better in tems of loss/metric convergence.

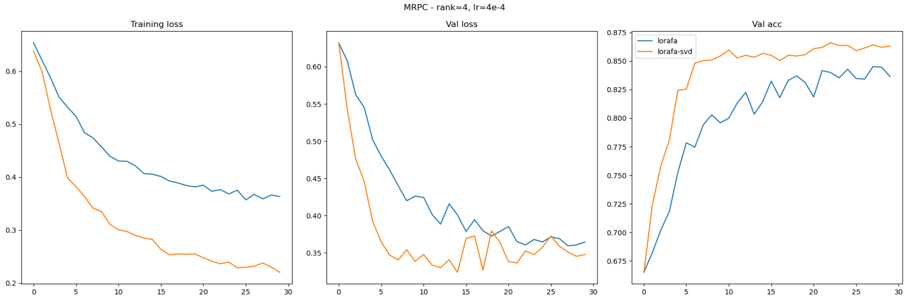
Skoltech

# MRPC

rank=4, learning rate=0.005

# MRPC

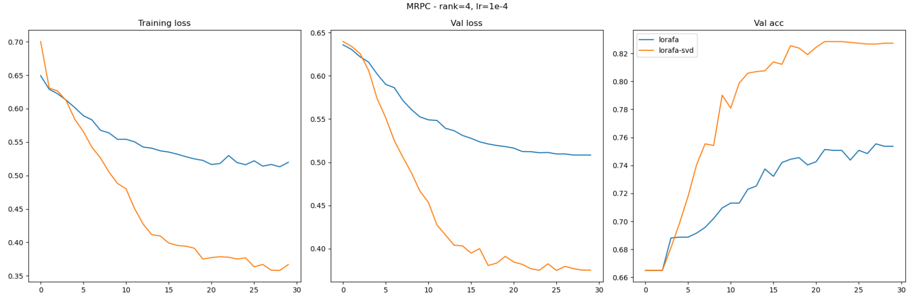
rank=4, learning rate=4e-4




Skoltech

# MRPC

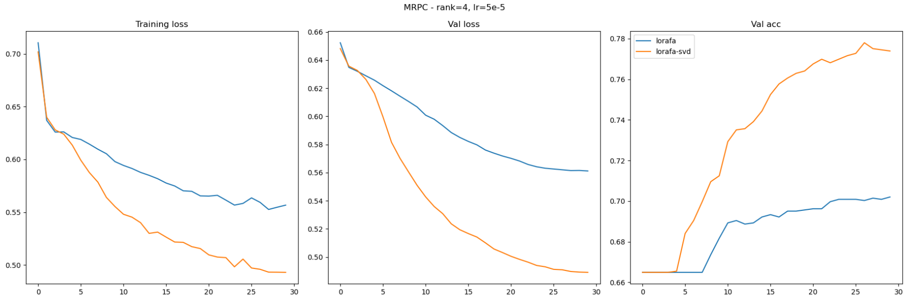
rank=4, learning rate=1e-4




Skoltech

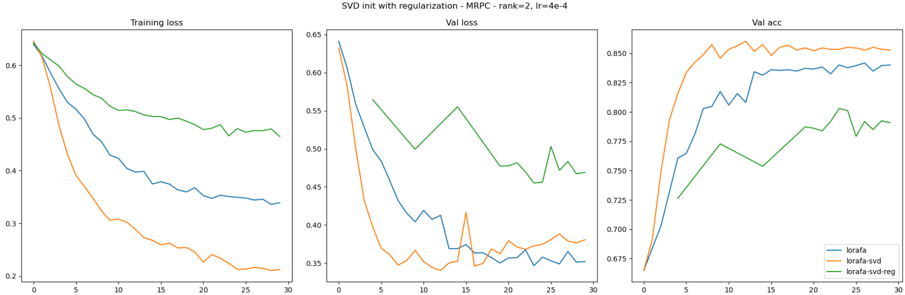
# MRPC

rank=4, learning rate=5e-5



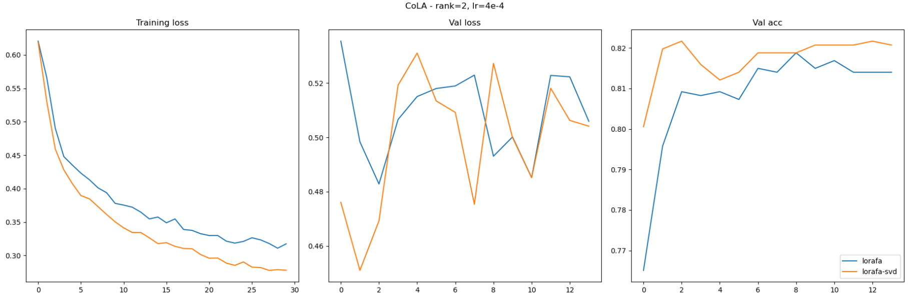
# MRPC - Regularization effect

Despite our hopes, regularization added to loss of form $\alpha \sum_{B \in lora\ \ layers} \|B - V\|_F^2$ performed much worse.



# COLA

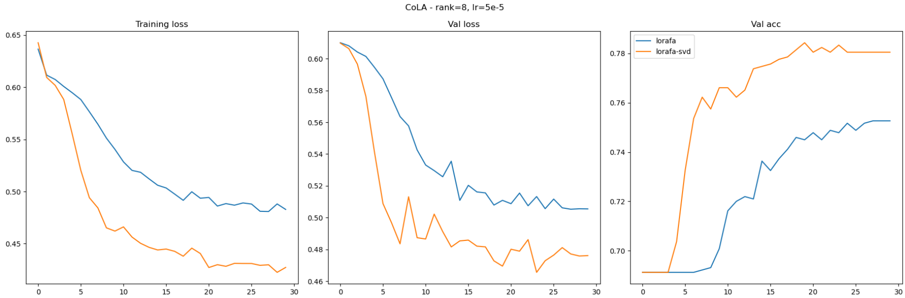
rank=2, learning rate=4e-4




Skoltech

# COLA

rank=8, learning rate=5e-5



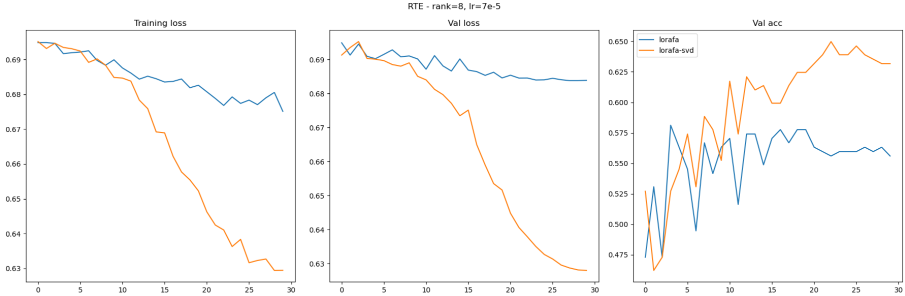
# RTE

rank=4, learning rate=5e-5


Skoltech

# RTE

rank=8, learning rate=7e-5




Skoltech

# RTE

rank=2, learning rate=4e-4

# SST-B

rank=2, learning rate=4e-4, metric = Pearson correlation


Skoltech

# Conclusions

- SVD initialization is indeed informative (performs same or better in all experiments, faster convergence)
- training with much smaller LR becomes feasible

Skoltech

# Future directions

- experiments with other models/hyperparameters
- compare with other LoRA variants
- theoretical analysis

Skoltech

# Contributions

- **Sergey Karpukhin** - lora backend, regularization experiments
- **Yulia Sergeeva** - experiments with COLA
- **Pavel Bartenev** - experiments with RTE
- **Pavel Tikhomirov** - experiment design and backend, experiments with MRPC
- **Maksim Komiakov** - experiments with STS-B

Skoltech

# Thank you for your attention!

https://github.com/shredder67/svd-lorafa

Skoltech