Finetuning with LoRA

gollnickdata.de

LoRA

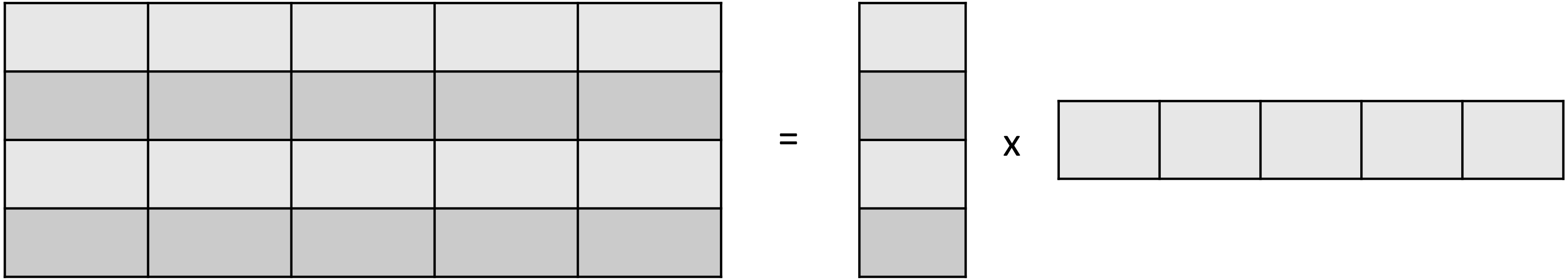
Introduction

- LoRA…Low-Rank Adaptation
- developed by Microsoft
- parameter-efficient fine-tuning (PEFT)
- only small subset of parameters updated instead of complete model
- reduces computational costs, and memory usage
- used technology -Low-Rank Decomposition (and quantisation)

LoRA

Decomposition

- matrix decomposition of weight matrix into trainable parameters
- number of parameters can be strongly reduced
- BUT exact result cannot be reached



LoRA

Reduction of Parameters

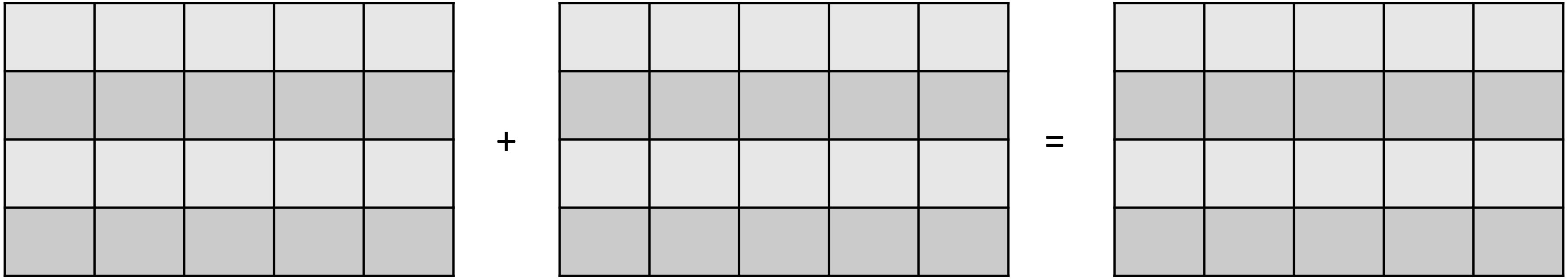
- matrix decomposition of weight matrix into change matrices
- number of parameters can be strongly reduced

Number of Parameters	**Matrix Dimensions**	**Rank 1 Nr. Parameters**
100	10 x 10	20
1 M	1000 x 1000	2000

LoRA

Reduction of Parameters

- matrix decomposition of weight matrix into change matrices
- number of parameters can be strongly reduced



LoRA Weights
(only changes)

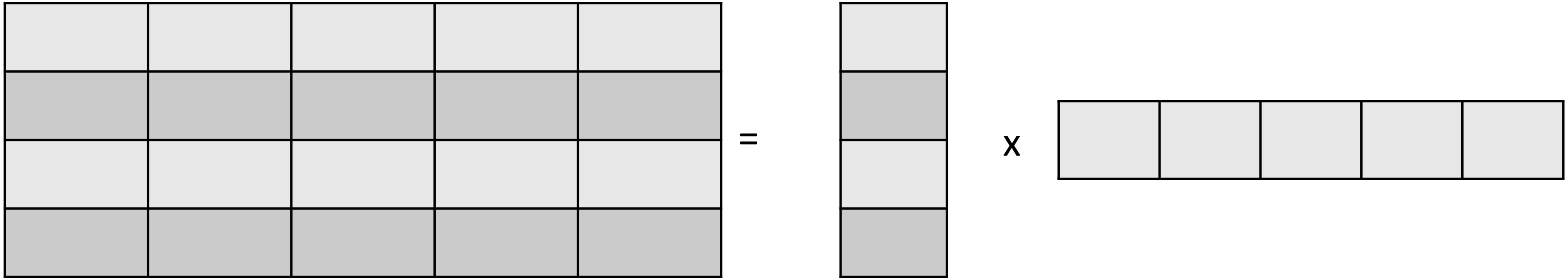
Original Model Weights

Fine-Tuned
Model Weights

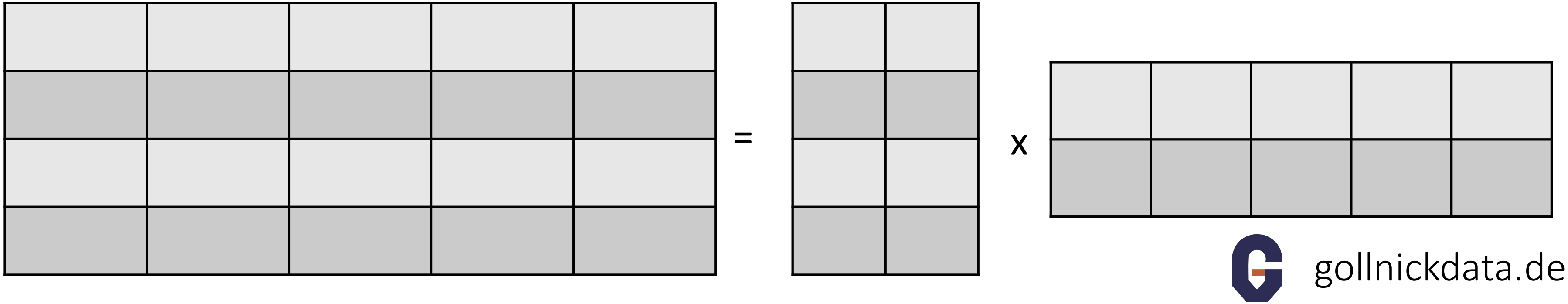
LoRA

Rank

Rank 1



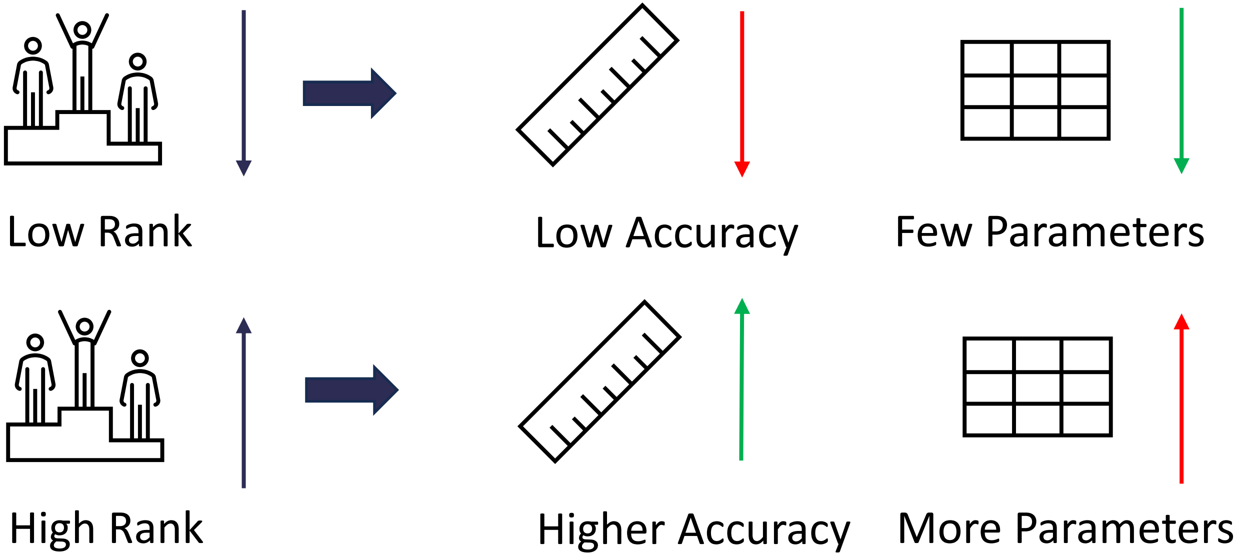
Rank 2



LoRA

Rank Selection

- Microsoft used Rank 8 to 16 in ist paper
- other sources use rank 32 or more



Low Rank → Low Accuracy, Few Parameters

High Rank → Higher Accuracy, More Parameters