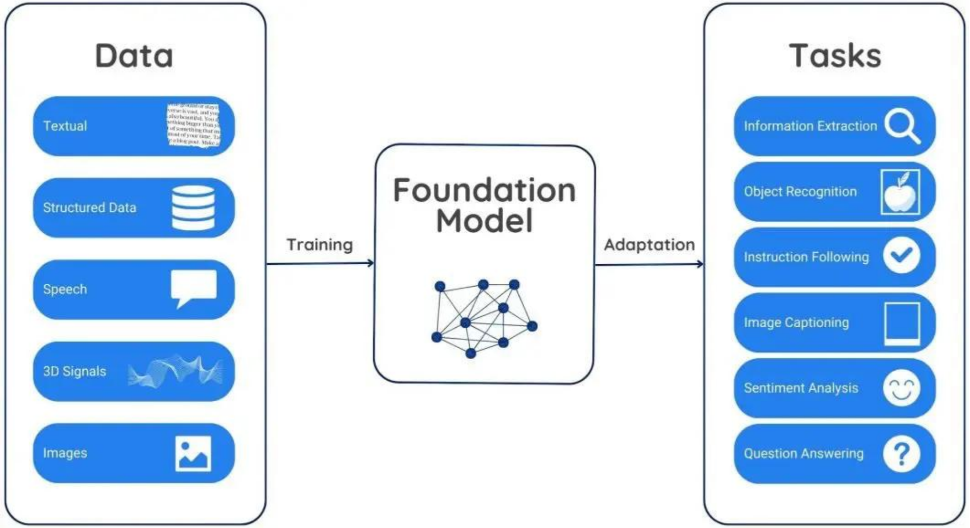
Parameter Efficient Fine-tuning via Explained Variance Adaptation

Fabian Paischer*, Lukas Hauzenberger*, Thomas Schmied, Benedikt Alkin, Marc Peter Deisenroth, Sepp Hochreiter

JOHANNES KEPLER UNIVERSITY LINZ

Problem Statement

- Foundation Models [1] are trained on vast collections of data
    - Adaptation of large pre-trained models to downstream tasks is expensive!

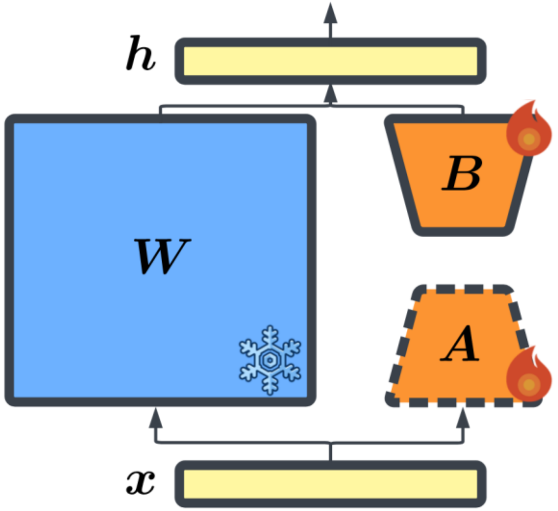


JOHANNES KEPLER UNIVERSITY LINZ

[1] On the Opportunities and Risks of Foundation Models, Bommasani et al., 2021

Background - LoRA

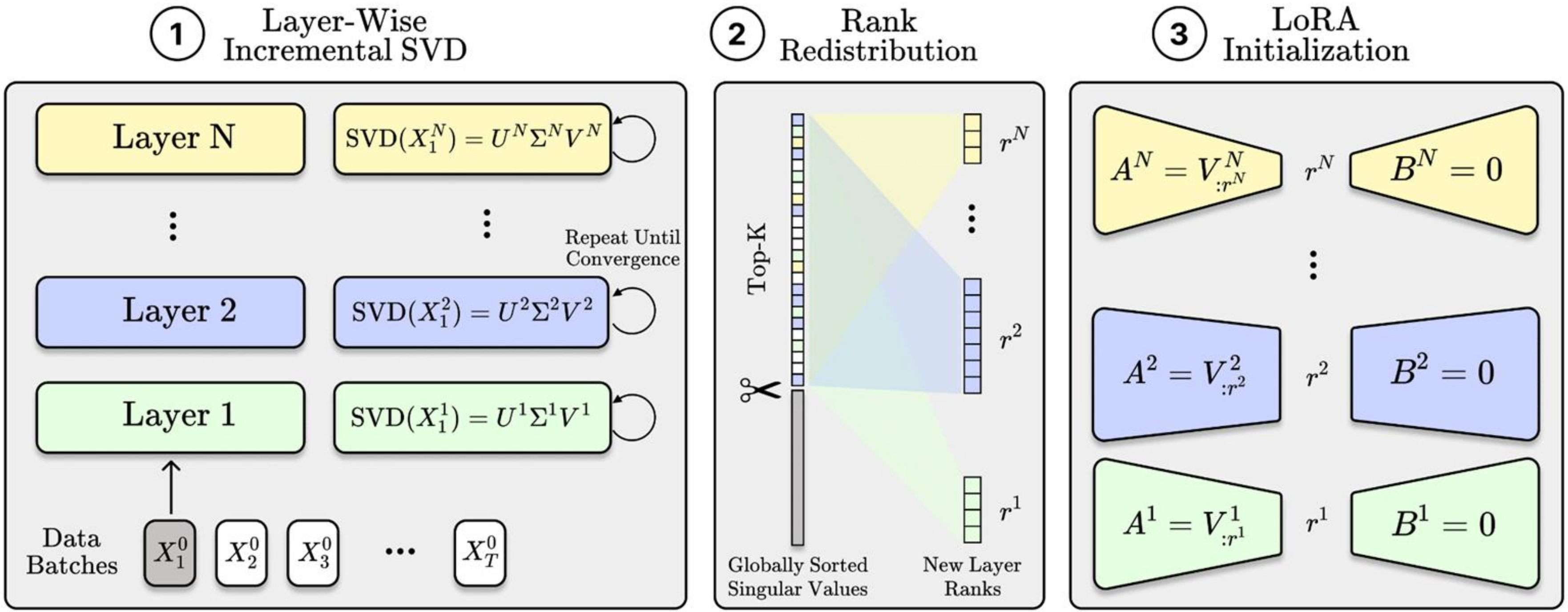
- LoRA (Hu et al., 2022) drastically reduces cost of finetuning
  - Introduce low rank weight matrices: W' = W + BA
  - Only fine-tunes newly introduced weights



- LoRA is usually initialized randomly with uniform ranks
  - How to best initialize LoRA?
  - How can we allocate ranks in task-driven manner?

JOHANNES KEPLER UNIVERSITY LINZ

Explained Variance Adaptation (EVA)



JKU JOHANNES KEPLER UNIVERSITY LINZ

LoRA Initialization Strategies

Method	Initialization	Adaptive ranks
LoRA (Hu et al., 2022)	Random	✗
AdaLoRA (Zhang et al., 2023a)	Random	✓
PiSSA (Meng et al., 2024)	Weight-driven	✗
MiLoRA (Wang et al., 2024a)	Weight-driven	✗
OLoRA (Büyükakyüz, 2024)	Weight-driven	✗
LoRA-GA (Wang et al., 2024b)	Data-driven	✗
CorDA (Yang et al., 2024)	Data-driven	✗
EVA (Ours)	Data-driven	✓

Initialization stage - Incremental SVD

Initialization	Method	Memory (GB)	% of Training
Weight-driven	PiSSA / MiLoRA	-	1.5
	OLoRA	-	0.1
Data-driven	LoRA-GA$_{bs=8}$	56.95	2.4
	CorDA$_{bs=1}$	55.64	4.5
	EVA$_{bs=16}$	32.85	0.7
	EVA$_{bs=8}$	29.39	0.3
	EVA$_{bs=4}$	27.51	0.2

Initialization time for Llama-2-7b on Common Sense reasoning datasets on a single A100

Experiments

1. Math and common-sense reasoning tasks
   a. Llama 2 7B
   b. Llama 3.1 8B
   c. Gemma 2 9B
   d. Gemma 2 27B *
   e. Llama 3.1 70B *

2. Language Understanding tasks
   a. DeBERTa v3 Base (184M)
   b. RoBERTa Large (355M)

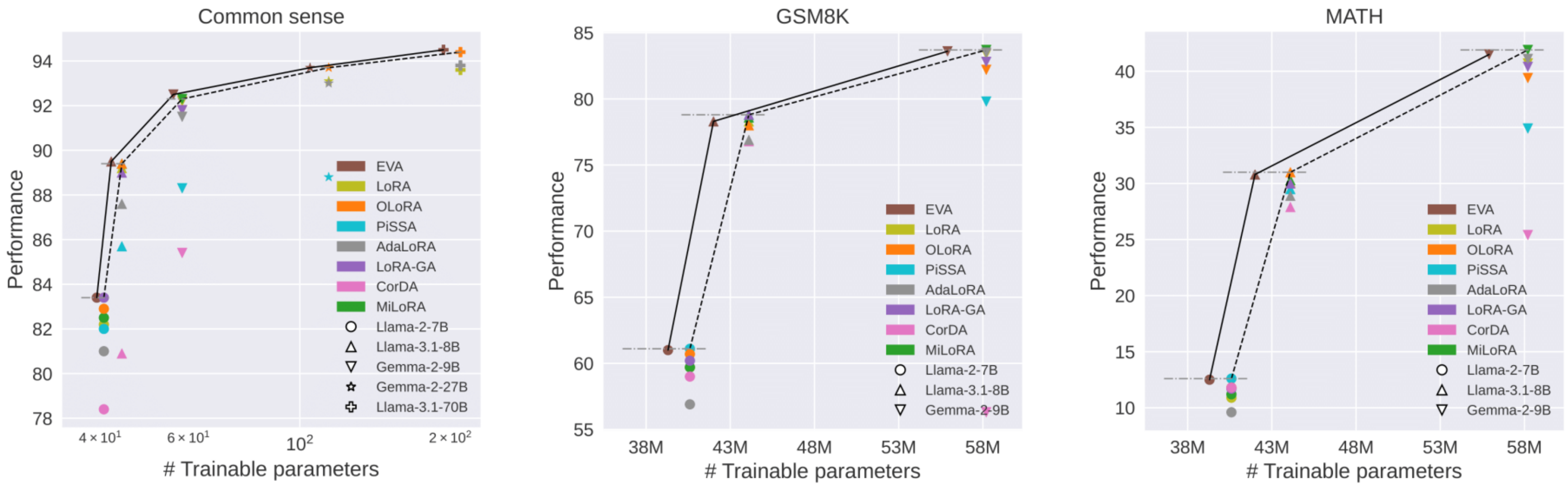
3. Image classification
   a. DINOv2-g/14 (1.1B)

4. Decision making
   a. Decision Transformers (12M)

* only common-sense reasoning

JOHANNES KEPLER UNIVERSITY LINZ

Results


Common sense
Performance
Trainable parameters
GSM8K
MATH
EVA
LoRA
OLoRA
PiSSA
AdaLoRA
LoRA-GA
CorDA
MiLoRA
Llama-2-7B
Llama-3.1-8B
Gemma-2-9B
Gemma-2-27B
Llama-3.1-70B
38M
43M
48M
53M
58M

Language Understanding

Model	Method	MNLI	QNLI	QQP	SST2	CoLA	MRPC	RTE	STS-B	Avg
RoBERTa Large	FFT	90.2	94.7	**92.2**	**96.4**	68.0	90.9	86.6	92.4	88.9
	LoRA	90.7$_{\pm.1}$	94.8$_{\pm.1}$	92.0$_{\pm.0}$	96.2$_{\pm.3}$	69.1$_{\pm.5}$	91.1$_{\pm.6}$	88.1$_{\pm1.1}$	92.3$_{\pm.1}$	89.3
	AdaLoRA	90.5$_{\pm.1}$	94.8$_{\pm.2}$	90.6$_{\pm.1}$	96.1$_{\pm.2}$	68.2$_{\pm.7}$	90.7$_{\pm.6}$	84.4$_{\pm.9}$	91.8$_{\pm.1}$	88.4
	PiSSA	90.1$_{\pm.1}$	94.7$_{\pm.0}$	91.0$_{\pm.0}$	96.1$_{\pm.2}$	68.7$_{\pm1.3}$	90.4$_{\pm.6}$	87.6$_{\pm.5}$	92.5$_{\pm.3}$	88.9
	OLoRA	**90.9**$_{\pm.1}$	**95.0**$_{\pm.1}$	92.0$_{\pm.2}$	96.3$_{\pm.3}$	69.0$_{\pm1.5}$	91.0$_{\pm1.0}$	87.9$_{\pm1.2}$	92.4$_{\pm.1}$	89.3
	LoRA-GA	90.8$_{\pm.2}$	94.9$_{\pm.1}$	92.0$_{\pm.0}$	96.3$_{\pm.4}$	68.4$_{\pm1.9}$	91.0$_{\pm.2}$	87.0$_{\pm.4}$	92.3$_{\pm.3}$	89.1
	CorDA	89.3$_{\pm.0}$	92.6$_{\pm.0}$	89.7$_{\pm.0}$	95.5$_{\pm.0}$	67.8$_{\pm1.0}$	90.1$_{\pm.9}$	86.5$_{\pm.8}$	91.8$_{\pm.2}$	87.9
	EVA	90.8$_{\pm.1}$	**95.0**$_{\pm.2}$	92.1$_{\pm.1}$	96.2$_{\pm.1}$	**69.5**$_{\pm1.4}$	**91.4**$_{\pm.8}$	**88.8**$_{\pm1.2}$	**92.6**$_{\pm.1}$	**89.6**
	DoRA	89.5$_{\pm.1}$	94.6$_{\pm.1}$	89.9$_{\pm.1}$	96.1$_{\pm.1}$	69.3$_{\pm.8}$	91.0$_{\pm.6}$	88.4$_{\pm1.2}$	92.4$_{\pm.1}$	88.9
DeBERTa v3 Base	FFT	90.1	94.0	92.4	95.6	69.2	89.5	83.8	91.6	88.3
	LoRA	90.5$_{\pm.1}$	94.3$_{\pm.1}$	92.4$_{\pm.1}$	95.2$_{\pm.3}$	72.0$_{\pm1.3}$	91.4$_{\pm.7}$	88.9$_{\pm.5}$	91.7$_{\pm.1}$	89.6
	AdaLoRA	**90.8**	**94.6**	92.2	96.1	71.5	90.7	88.1	91.8	89.5
	PiSSA	90.1$_{\pm.3}$	94.1$_{\pm.1}$	91.8$_{\pm.1}$	95.8$_{\pm.1}$	**72.7**$_{\pm1.7}$	90.9$_{\pm.6}$	86.5$_{\pm1.2}$	91.6$_{\pm.2}$	89.2
	OLoRA	90.5$_{\pm.1}$	94.4$_{\pm.1}$	**92.6**$_{\pm.1}$	**96.2**$_{\pm.2}$	72.0$_{\pm1.0}$	91.6$_{\pm.7}$	89.1$_{\pm.9}$	**92.0**$_{\pm.2}$	89.8
	LoRA-GA	89.8$_{\pm.7}$	**94.6**$_{\pm.1}$	92.2$_{\pm.0}$	95.6$_{\pm.8}$	72.2$_{\pm.9}$	90.8$_{\pm.9}$	86.6$_{\pm1.1}$	90.5$_{\pm.6}$	89.0
	CorDA	90.0$_{\pm.1}$	93.8$_{\pm.1}$	91.1$_{\pm.1}$	95.5$_{\pm.4}$	71.8$_{\pm1.2}$	89.6$_{\pm.5}$	83.9$_{\pm.3}$	91.1$_{\pm.2}$	88.3
	EVA	90.6$_{\pm.1}$	94.4$_{\pm.1}$	92.4$_{\pm.0}$	**96.2**$_{\pm.2}$	72.5$_{\pm1.3}$	91.8$_{\pm.6}$	**89.4**$_{\pm.7}$	**92.0**$_{\pm.2}$	**89.9**
	DoRA	89.0$_{\pm.2}$	94.1$_{\pm.1}$	88.0$_{\pm.1}$	94.6$_{\pm.4}$	70.3$_{\pm.5}$	**91.9**$_{\pm.6}$	87.8$_{\pm.7}$	91.8$_{\pm.1}$	88.4

JOHANNES KEPLER UNIVERSITY LINZ

Try out EVA

```
from peft import EvaConfig, LoraConfig, get_peft_model, initialize_lora_eva_weights

eva_config = EvaConfig(
    rho=rho
)
peft_config = LoraConfig(
    init_lora_weights="eva",
    eva_config=eva_config
)

peft_model = get_peft_model(model, peft_config, low_cpu_mem_usage=True)
initialize_lora_eva_weights(peft_model, dataloader)
```

Conclusion

- EVA leverages downstream data to initialize low rank weights and adaptively allocates ranks

- EVA reaches highest average score with less trainable parameters on
  - Language generation
  - Language understanding
  - Image classification
  - Decision making

PEFT
SCAN ME

Code
SCAN ME

Paper
SCAN ME

JOHANNES KEPLER UNIVERSITY LINZ

ellis
European Laboratory for Learning and Intelligent Systems

References

[1] On the Opportunities and Risks of Foundation Models, Bommasani et al., 2021
[2] LoRA: Low-Rank Adaptation of Large Language Models, Hu et al., ICLR 2022
[3] Llama 2: Open Foundation and Fine-Tuned Chat Models, Touvron et al., arXiv:2307.09288
[4] The Llama 3 Herd of Mdoels, Llama Team, AI @ Meta, arXiv:2407.21783
[5] Gemma 2: Improving Open Language Models at a Practical Size, Gemma Team, arXiv:2408.00118
[6] DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, He et al., ICLR 2023
[7] RoBERTa: A Robustly Optimized BERT Pretraining Approach, Liu et al., arXiv:1907.11692
[8] DINOv2: Learning Robust Visual Features without Supervision, Oquab et al., TMLR 2024
[9] Decision Transformer: Reinforcement Learning via Sequence Modeling, Chen et al., NeurIPS 2021
[10] DoRA: Weight-Decomposed Low-Rank Adaptation, Liu et al., ICML 2024
[11] Training Verifiers to Solve Math Word Problems, Cobbe et al., arXiv:2110.14168
[12] Measuring Mathematical Problem Solving With the MATH Dataset, Hendrycks et al., NeurIPS 2021
[13] GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, Wang et al., ICLR 2019
[14] MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models, Yu et al., ICLR 2024

JOHANNES KEPLER
UNIVERSITY LINZ

ellis
European Laboratory for Learning and Intelligent Systems