

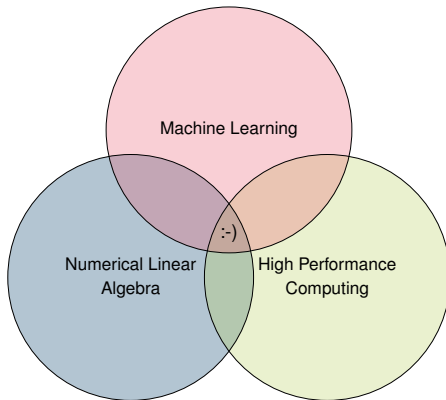


openGPT-X

Mathematical Techniques to Reduce Memory Requirements in Deep Learning

November 5, 2024 | Dr. Carolin Penke | Jülich Supercomputing Centre

MY RESEARCH INTERESTS



TRAINING LARGE MODELS

Training these large models needs

- Lots of computational resources (GPUs!),
- Lots of data.

Pretraining happens on supercomputers.



(R-U. Limbach / Forschungszentrum Jülich)

Finetuning of smaller models happens on workstations.

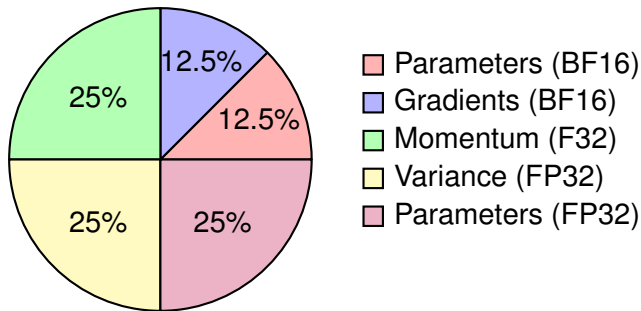


NVIDIA

In both settings, you want to use limited resources efficiently.

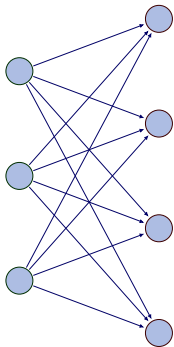
GPU MEMORY REQUIREMENTS DURING TRAINING

Using the mixed-precision Adam optimizer.



- + Activations, depending on sequence length and batch size.
- Activations can be reduced using activation checkpointing.

MATRICES EVERYWHERE



Parameter matrix

0.23	-0.15	0.5
0.1	0.45	-0.35
-0.2	0.3	0.25
0.4	-0.1	-0.05

Gradient matrix

-0.05	0.2	-0.1
0.15	-0.25	0.05
0.1	-0.15	0.3
-0.05	0.4	-0.2

Momentum matrix

0.01	0.02	-0.01
-0.03	0.04	-0.02
0.05	-0.01	0.06
-0.04	0.03	-0.05

Variance matrix

0.1	0.15	0.2
0.05	0.12	0.18
0.22	0.25	0.3
0.08	0.1	0.13

A layer in a neural network is represented by matrices.

LOW-RANK APPROXIMATIONS

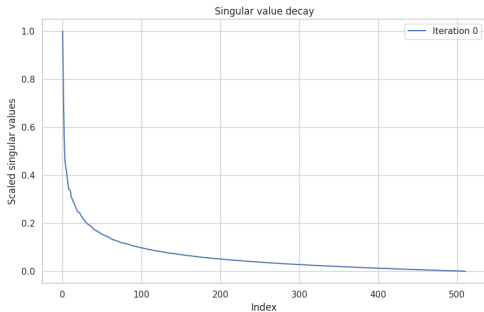
- When a matrix has (numerical) low rank, it can be approximated well by smaller matrices.

$$\begin{array}{ccc} \begin{array}{|c|} \hline \mathbf{G} \\ \hline \end{array} & \approx & \begin{array}{|c|} \hline \mathbf{L} \\ \hline \end{array} \times \begin{array}{|c|} \hline \mathbf{R}^T \\ \hline \end{array} \\ m \times n & & m \times k \qquad k \times n \end{array}$$

- Numerical low rank can be observed for **gradients**, momentum and variance.
→ These matrices can be compressed.

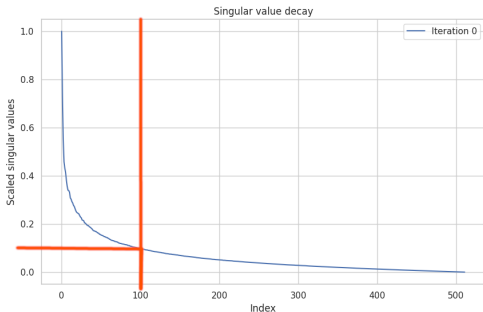
OBSERVING LOW RANK

The singular values of a matrix describe, how well a matrix can be approximated with a low-rank decomposition.



OBSERVING LOW RANK

The singular values of a matrix describe, how well a matrix can be approximated with a low-rank decomposition.



Here, a low rank decomposition with $k = 100$ (instead of $n = 512$) has an approximation quality of 90%.

OBSERVING LOW RANK

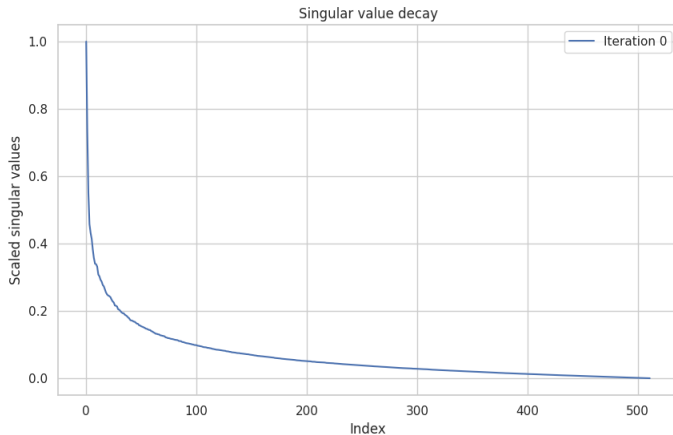


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

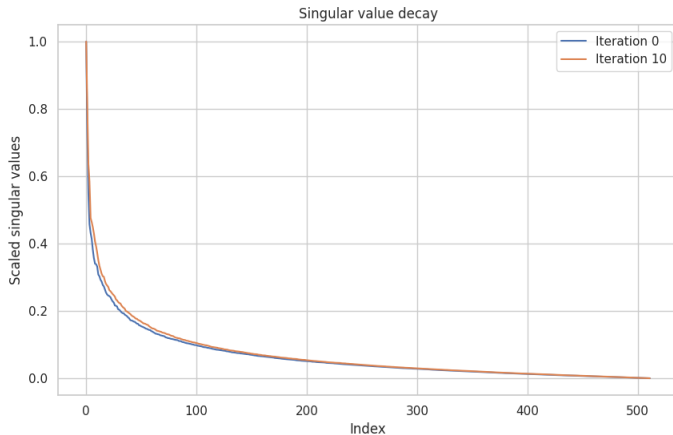


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

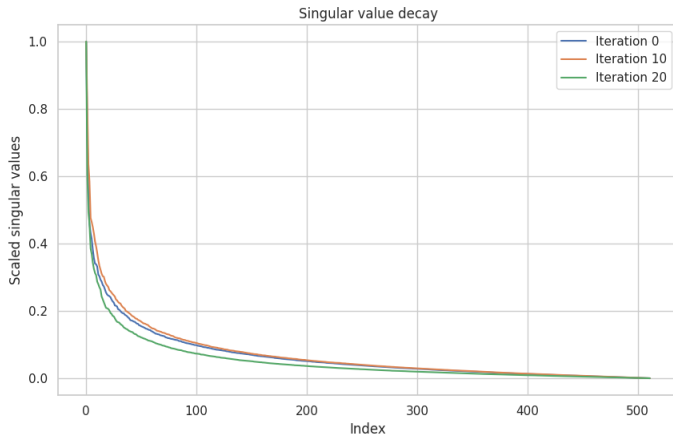


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

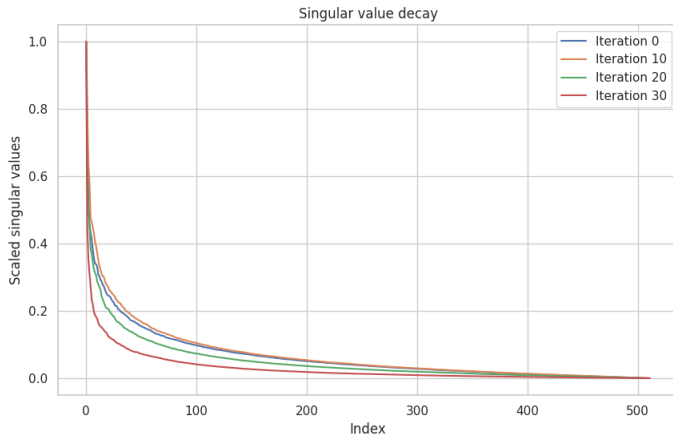


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

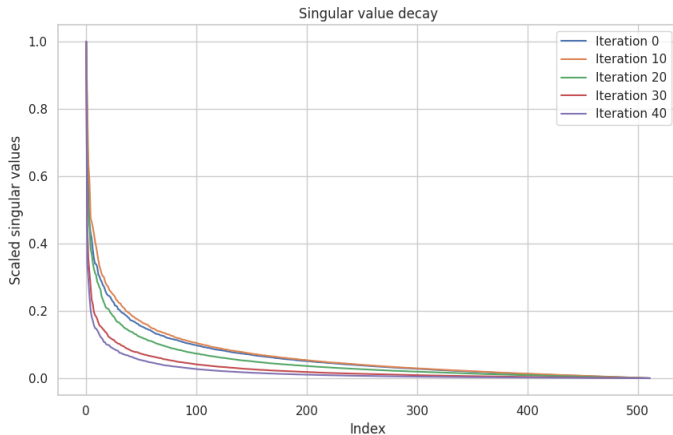


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

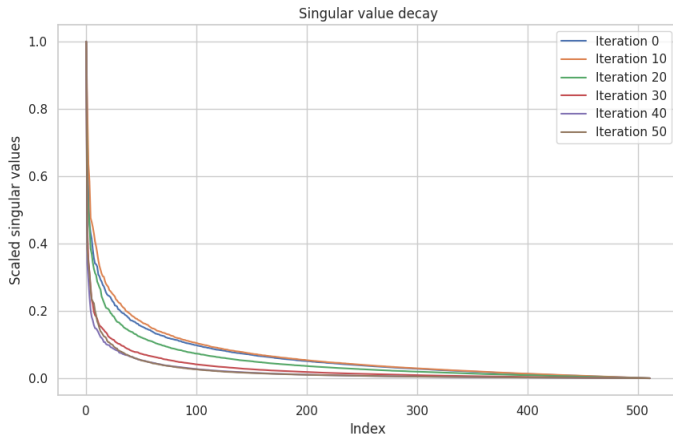


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

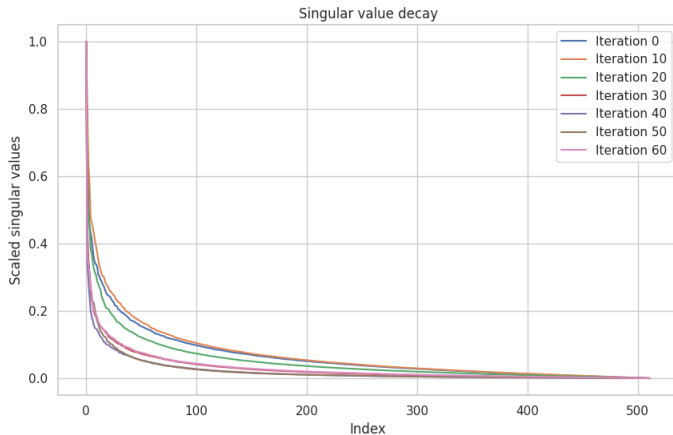


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

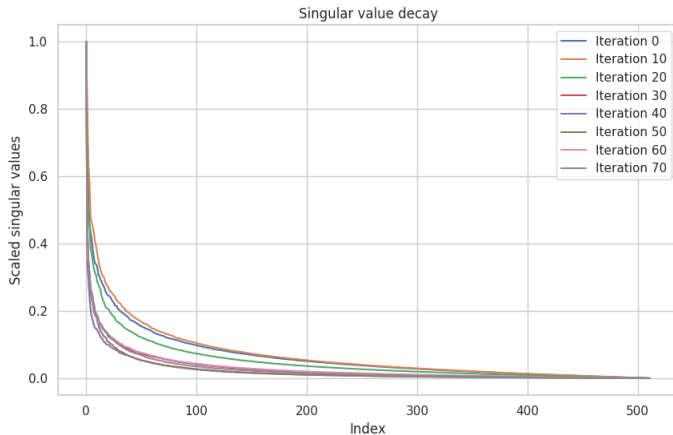


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

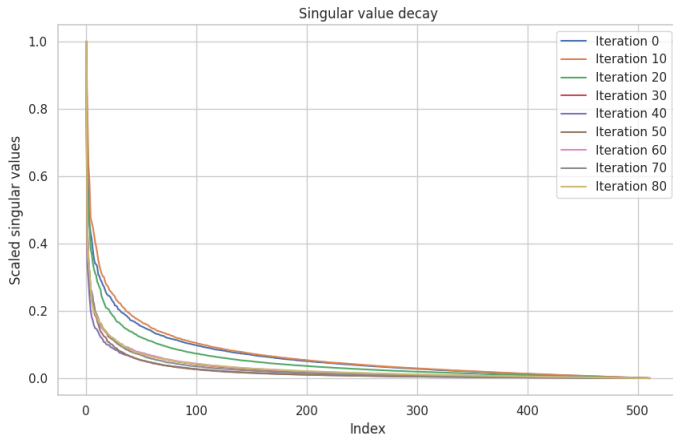


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

OBSERVING LOW RANK

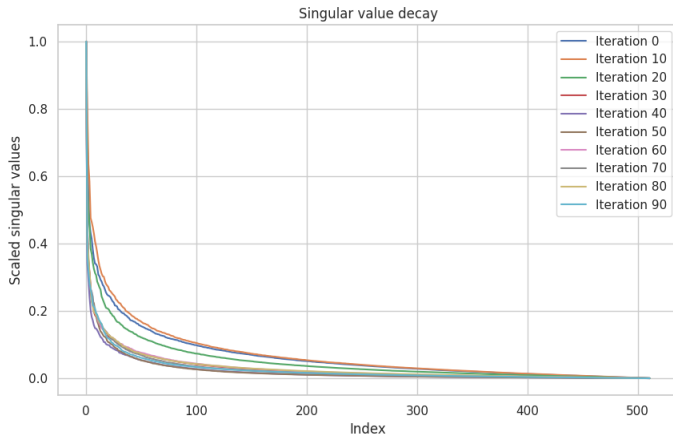


Figure: Singular value decay of gradient for first layer in pre-training 60M Llama model after various iterations.

EXPLOITING LOW RANK

LoRA: Low-Rank Adaptation of Large Language Models

- Established method for finetuning LLMs under memory constraints.

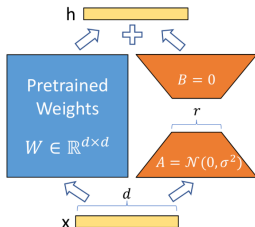
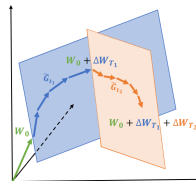


Figure 1: Our reparametrization. We only train A and B .

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "LoRA: Low-Rank Adaptation of Large Language Models", 2021.

GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection

- New method, projecting to low-rank subspace computed from gradient matrix.
- Lower memory footprint, better suited for pre-training.



J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, Y. Tian. "GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection", 2024.

SINGULAR VALUE DECOMPOSITION IN GALORE

Algorithm 2: Adam with GaLore

Input: A layer weight matrix $W \in \mathbb{R}^{m \times n}$ with $m \leq n$. Step size η , scale factor α , decay rates β_1, β_2 , rank r , subspace change frequency T .

Initialize first-order moment $M_0 \in \mathbb{R}^{n \times r} \leftarrow 0$

Initialize second-order moment $V_0 \in \mathbb{R}^{n \times r} \leftarrow 0$

Initialize step $t \leftarrow 0$

repeat

$G_t \in \mathbb{R}^{m \times n} \leftarrow -\nabla_W \varphi_t(W_t)$

if $t \bmod T = 0$ **then**

$U, S, V \leftarrow \text{SVD}(G_t)$

$P_t \leftarrow U[:, :r]$ {Initialize left projector as $m \leq n$ }

else

$P_t \leftarrow P_{t-1}$ {Reuse the previous projector}

end if

$R_t \leftarrow P_t^\top G_t$ {Project gradient into compact space}

UPDATE(R_t) **by Adam**

$M_t \leftarrow \beta_1 \cdot M_{t-1} + (1 - \beta_1) \cdot R_t$

$V_t \leftarrow \beta_2 \cdot V_{t-1} + (1 - \beta_2) \cdot R_t^2$

$M_t \leftarrow M_t / (1 - \beta_1^t)$

$V_t \leftarrow V_t / (1 - \beta_2^t)$

$N_t \leftarrow M_t / (\sqrt{V_t} + \epsilon)$

$\tilde{G}_t \leftarrow \alpha \cdot P N_t$ {Project back to original space}

$W_t \leftarrow W_{t-1} + \eta \cdot \tilde{G}_t$

$t \leftarrow t + 1$

until convergence criteria met

return W_t

- GaLore computes a singular value decomposition to get subspace basis.
- The randomized range finder is a more efficient method.

J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, Y. Tian. “GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection”, 2024.

THE RANDOMIZED RANGE FINDER

The right tool for the job

ALGORITHM 4.1: RANDOMIZED RANGE FINDER

Given an $m \times n$ matrix \mathbf{A} , and an integer ℓ , this scheme computes an $m \times \ell$ orthonormal matrix \mathbf{Q} whose range approximates the range of \mathbf{A} .

- 1 Draw an $n \times \ell$ Gaussian random matrix $\mathbf{\Omega}$.
- 2 Form the $m \times \ell$ matrix $\mathbf{Y} = \mathbf{A}\mathbf{\Omega}$.
- 3 Construct an $m \times \ell$ matrix \mathbf{Q} whose columns form an orthonormal basis for the range of \mathbf{Y} , e.g., using the QR factorization $\mathbf{Y} = \mathbf{Q}\mathbf{R}$.

N. Halko, P.-G. Martinsson, J. A. Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions", 2010.

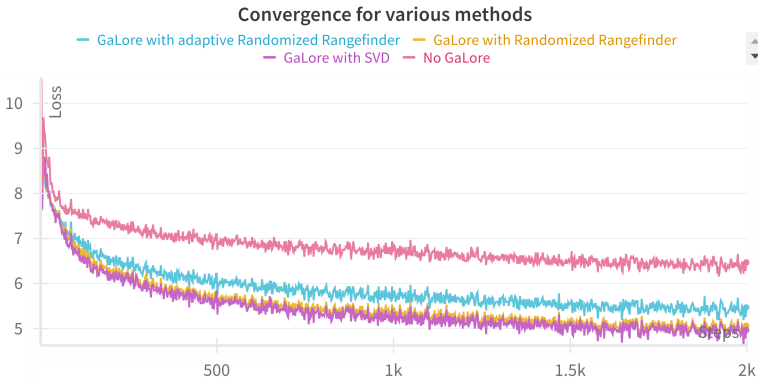
For an oversampling parameter $p \in \mathbb{N}$, $0 \leq p \leq r$, we have

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^T\mathbf{A}\|_2 \leq \left(1 + 11\sqrt{r} \cdot \sqrt{\min\{m, n\}}\right) \sigma_{r-p+1}$$

with a probability of at least $1 - 6 \cdot p^{-p}$ under mild assumptions on p .

PRELIMINARY RESULTS

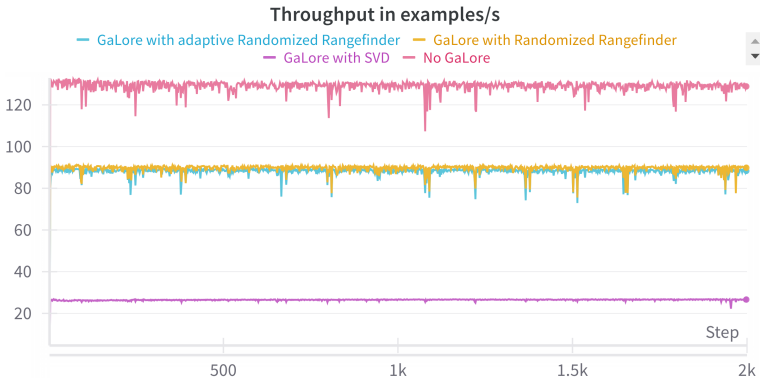
- Training a 60M Llama model, using rank 128, subspace computation in every step.



- Currently in development: GPU-optimized randomized rangefinder.

PRELIMINARY RESULTS

- Training a 60M Llama model, using rank 128, subspace computation in every step.







- Currently in development: GPU-optimized randomized rangefinder.

CONCLUSIONS

- OpenGPT-X spawned promising research directions in the area of efficient resource utilization during pre-training (AP1).
- More interesting research: Find [Chelsea John](#)'s poster.

Performance and Power: Systematic Evaluation of AI Workloads on Accelerators with CARAML

Chelsea Maria John , Stepan Nassyr , Carolin Penke , Andreas Herten 
Jülich Supercomputing Centre
Forschungszentrum Jülich
Jülich, Germany

Thank you for your attention!

This work was funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project OpenGPT-X (project no. 68GX21007D).