Low Rank Approximation of Generative Transformers

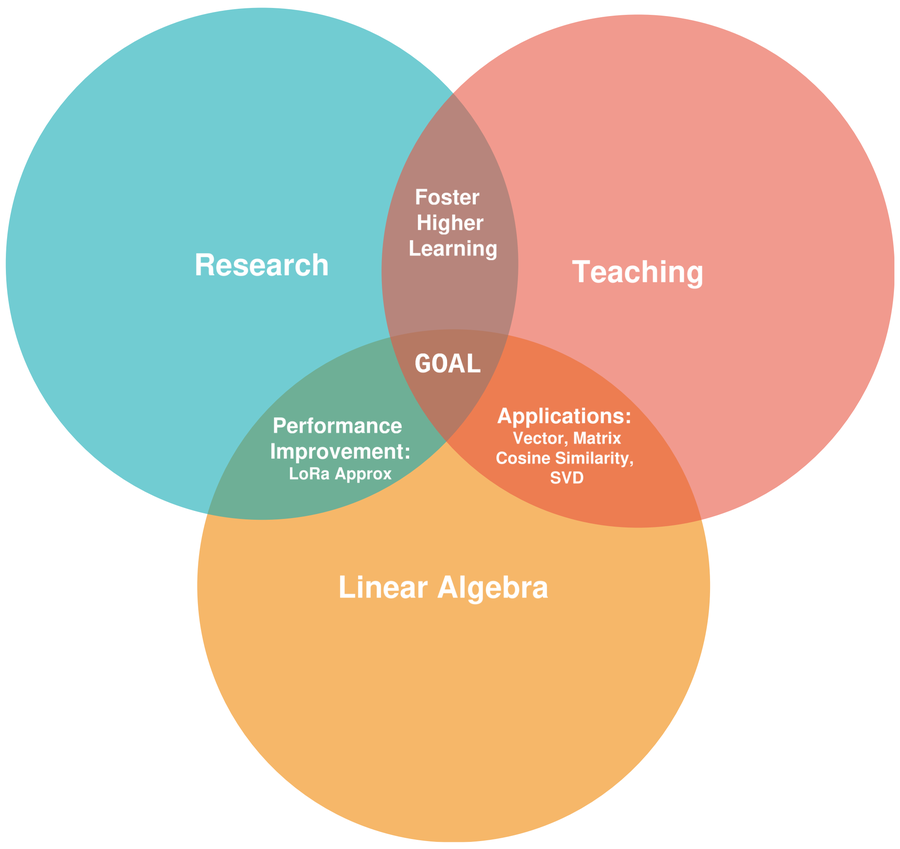
Using LLMs in Research and Education of Linear Algebra

Asger Song Poulsen and Hugo Daniel Macedo
Department of Electrical and Computer Engineering

Abstract*: Generative transformers are the foundational models changed our reality. The traditional approach to research and education assimilated generative AI and the area is redefining the fundamental skills we base our practice on. Linear algebra, a fundamental mathematical framework supporting data science, machine learning and engineering is now the engine of AI that future "prompt engineers" must understand. This project proposes/features a synergy of three domains: Research, Teaching, and Linear Algebra. We apply concepts taught in the classroom by showing its usage to build Generative AI (e.g.: self-attention matrices), we do research on how to improve such model performances by low rank approximation (singular-value decomposition), we foresee the usage of generative AI to aid students achieve higher levels of mastery over the linear algebra theory.*

Goal: Explore the usage of transformer models and generative AI in practice:

- Apply linear algebra concepts
- Motivate deeper study
- Advance performance

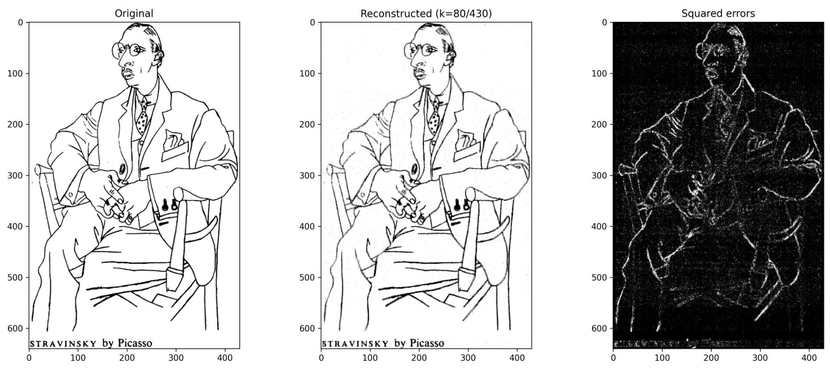


SW2PLA: A course of Linear algebra for software developers:

- Course aligns with students' future career
- Tasks on right domain: SVD to compress images
- Two level grading:
  - Use an API for LA for max 7 out of 12
  - Learn LA in depth for full grade
- A lab for GAI experiments:
  - Dedicated tutoring
  - Fine-tune LLMs using SVD
  - Does it foster deeper understanding?

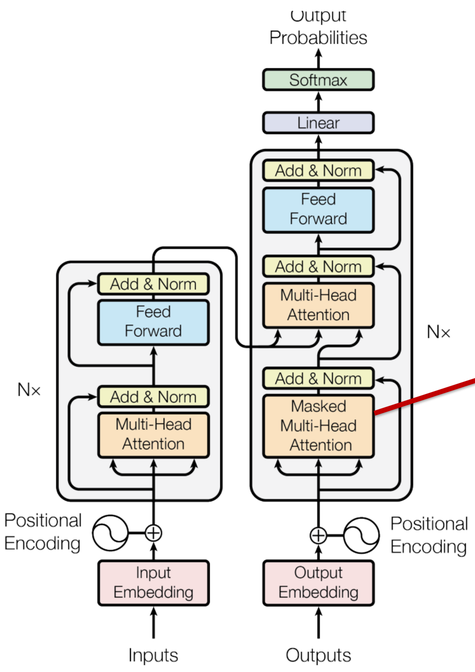
SVD: Singular Value Decomposition factorizes a matrix into three matrices :

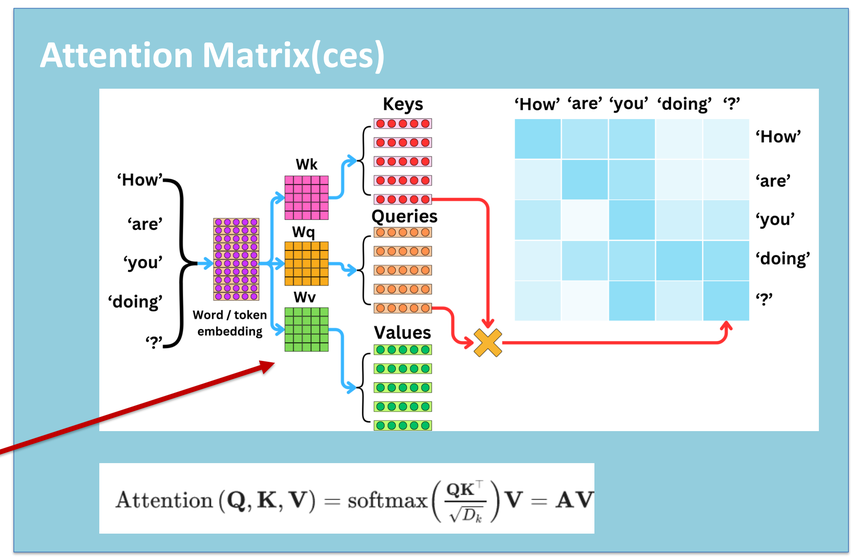
- left/right singular vectors and values: $A = U\Sigma V^T$
- Original as a sum of layers: $A = \sum_{i=1}^{rank} u_i \sigma_i v_i^T$



Transformers: A deep learning (Vaswani et al 2017) architecture

- Text is embedded into vectors
- Pre-trained with billions of parameters
- Accuracy is given by self-attention – a mechanism to focus on the relevant aspects of the text using matrices $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$



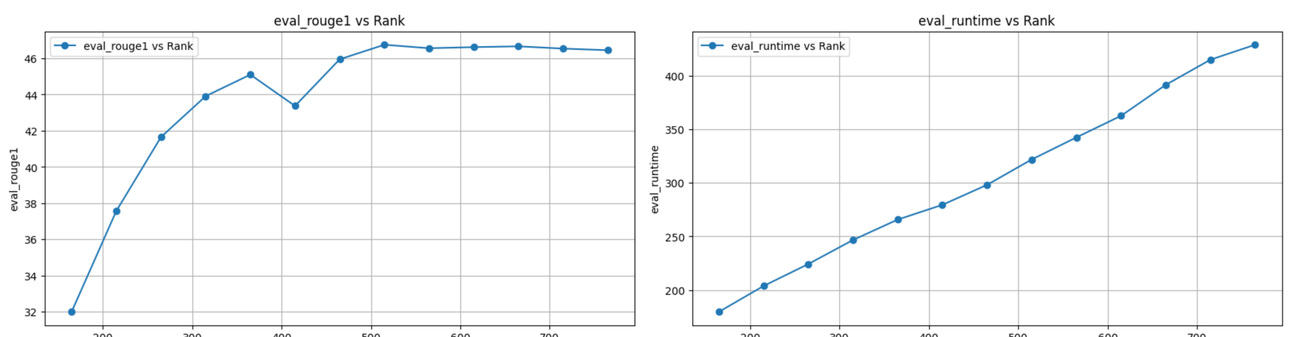


Low Rank Approximation of Attention:
Studies have shown Q, K and V matrices are good candidates for **SVD** factorization after training (Hu et al 2021).

$$\begin{matrix} Q \\ V \\ K \end{matrix} \xrightarrow[\text{SVD}]{} \begin{matrix} U_q\ \Sigma_q\ V_q^T \\ U_v\ \Sigma_v\ V_v^T \\ U_k\ \Sigma_k\ V_k^T \end{matrix}$$

Results: Our work on BART confirms that by performing **SVD** and cutting singular values:

- Computational performance improves and quality is maintained
- Enables a research quest on the right rank for each model/dataset



Example: Summarization task degradation:

Higher rank *closer to* BART

Finding the **right rank**

Lower rank *further from* BART

(less layers of the original model)

```
Rank 765: Hannah doesn't know Betty's number. She texted Larry last time they were at the park together.
Rank 660: Hannah doesn't know Betty's number. She texted Larry last time they were at the park together.
Rank 510: Hannah can't find Betty's number. She texted Larry last time they were at the park.
Rank 410: Hannah is looking for Betty's number. Amanda can't find it.
Rank 370: Amanda can't find Betty's number. Hannah texted Larry last time they were at the park.
Rank 365: Amanda is looking for Betty's number. Hannah doesn't know her boyfriend Larry.
Rank 360: Amanda can't find Betty's number. Hannah and Amanda don't know each other well.
Rank 315: Amanda and Hannah don't know each other.
Rank 210: Amanda and Amanda don't know what to do with their relationship with Larry.
Rank 190: Amanda and Amanda don't know each other person in the world.
Rank 165: Amanda and Amanda are going to the park.
```

References

- Vaswani, Ashish et al. (2017). *Attention Is All You Need*
- Hu, Edward J. et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*
- Jeff Bussgang (2024). *An AI Professor at Harvard: ChatLTV*