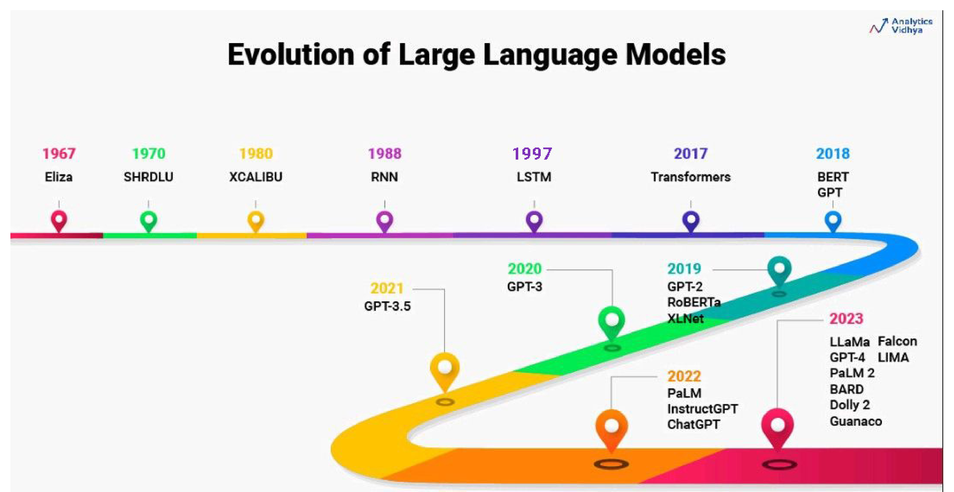
Towards Small Large Language Models

Nelie Laura MAKENNE, Prof. Ashwin ITTOO

HEC Liège/ULiège

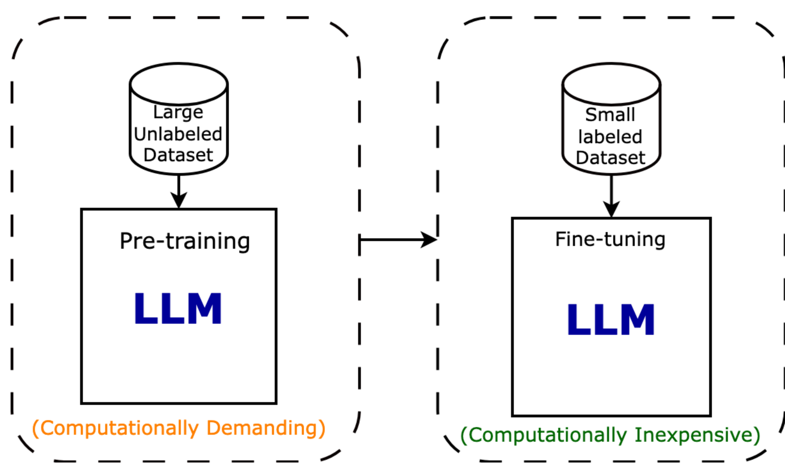
INTRODUCTION

Large Language Model (LLM) are advanced artificial intelligence designed to understand and generate human-like text based on massive amounts of data.



Their emergence has revolutionized recently the approach to natural language tasks by achieving state-of-the-art performance across various applications. Despite their remarkable performance, the size of LLMs poses significant challenges for their usage in lower-resource environments. Hence, our study will focus on addressing the following question:

Which techniques can we use to reduce the size of LLMs while maintaining essential information to minimize their computational and memory footprint?



Through a series experiments and evaluation, we aim to identify the most efficient methods for minimizing the computational and memory footprint of LLMs without compromising their performance on important tasks.

PROBLEM STATEMENT

Our problem centers of the resolution of the following optimization problem formulated as follows:

$$\min_k \mathcal{L}(W_k, W), \quad \text{subject to the constraint} \quad W_k = f(W), \tag{1}$$

where:

- $\mathcal{L}$ is the loss function;
- W_0 represents the pre-training weight matrix;
- W represents the updated weight matrix in the lower-dimensional space;
- f denotes the transformation function.

The aim is to find the optimal transformation function f that minimizes the loss function.

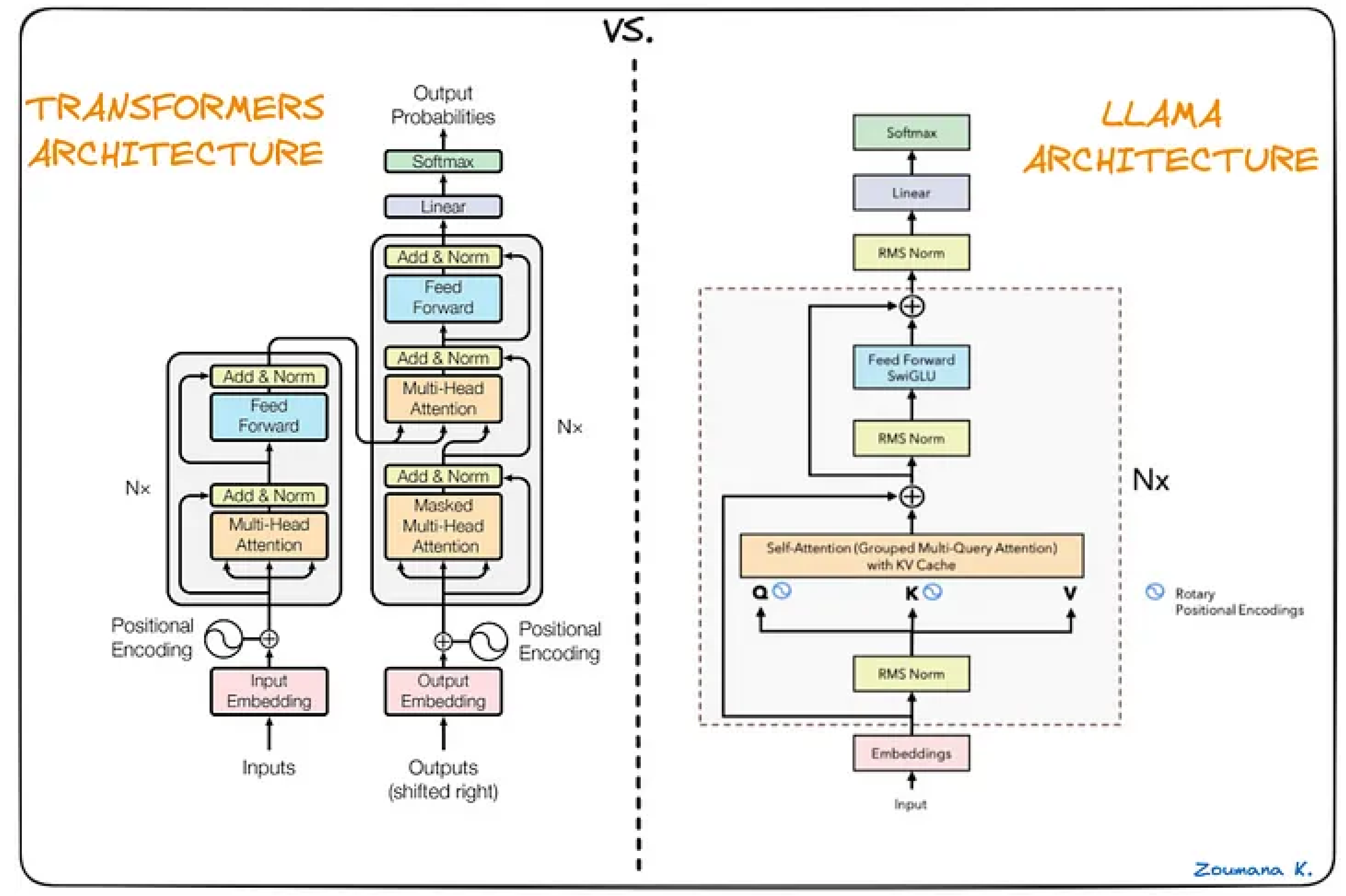
METHODOLOGY

Our methodology involves updating the parameters of the pre-trained language model (LLM) based on its new representation in a lower-dimensional space while minimizing information loss. While our method is yet to be implemented, we plan to adapt it for decoder-only transformer-based large language models, including:

- Llama 2 x Billion of parameters;
- Mixtral x Billion of parameters;
- Falcon x Billion of parameters.

Our goal is to optimize their performance to operate efficiently in resource-constrained settings while maintaining high accuracy and reliability through rigorous testing and evaluation, across various NLP tasks, including

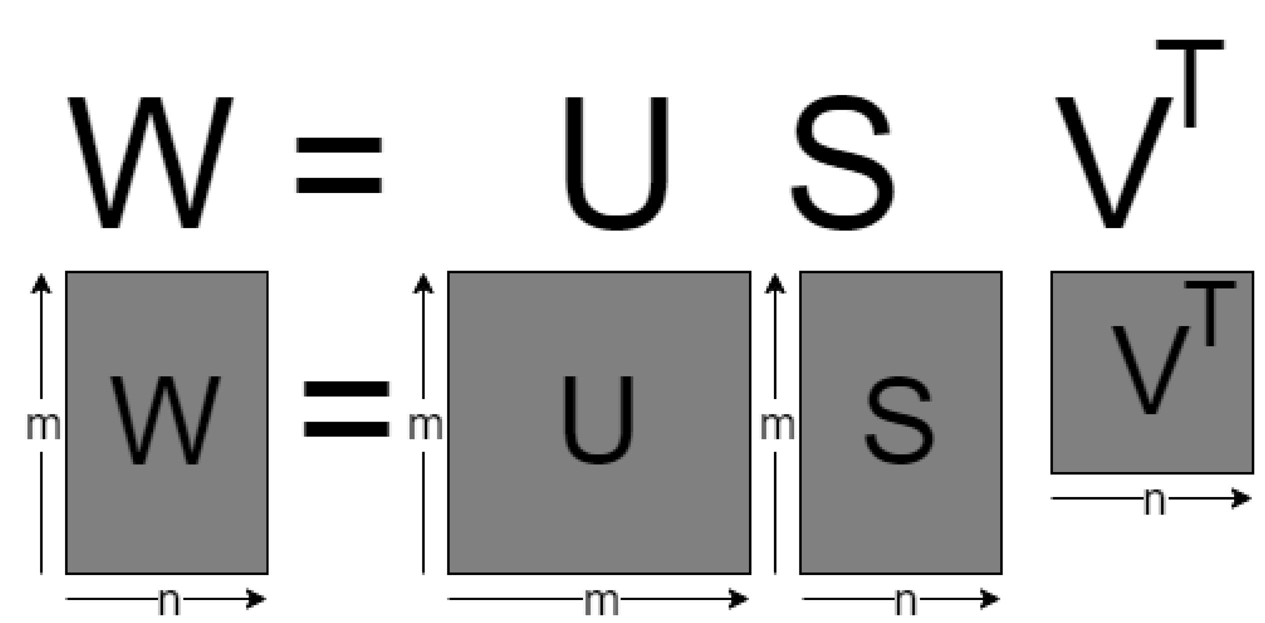
- Question-Answering, Intent Detection, and Topic Modeling



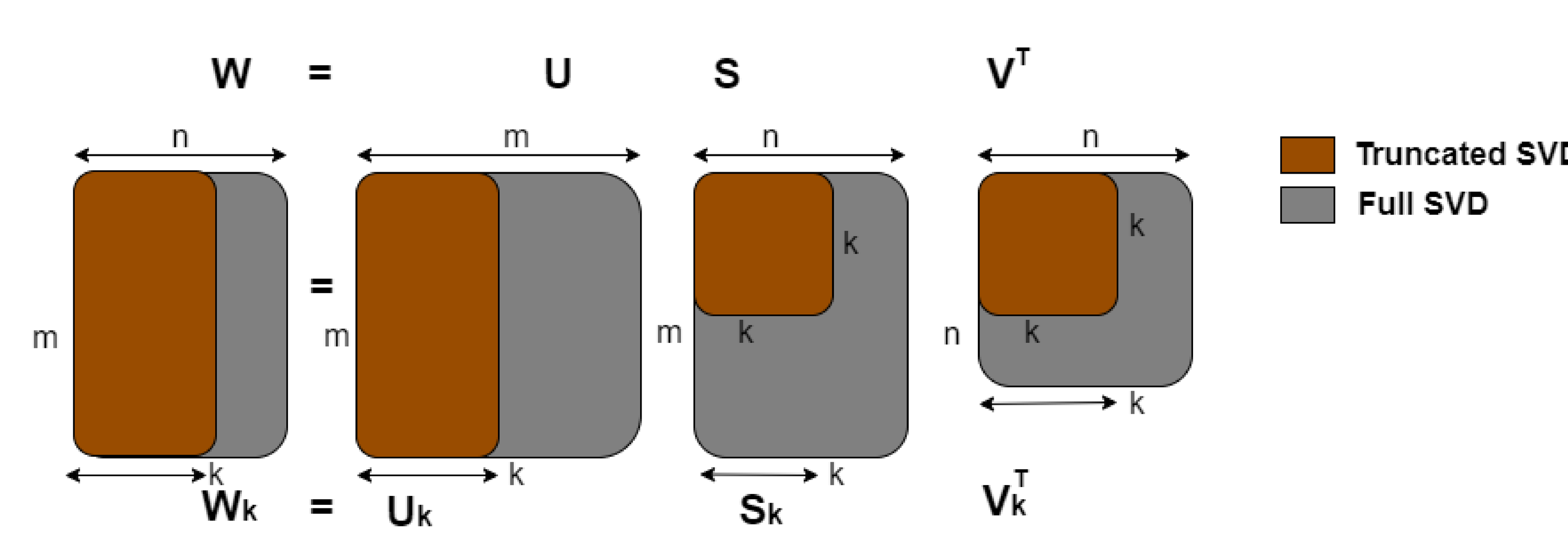
Our approach to finding the optimal transformation function involves implementing two dimensionality reduction techniques: **Singular Value Decomposition** (SVD), a classical mathematical method, and **Autoencoders**, a deep learning approach.

1. Singular Value Decomposition

Singular Value Decomposition is a matrix factorization method that decomposes a given matrix W into three matrices: U a left singular matrix, S a diagonal matrix containing singular values, and V a right singular matrix. U and V are orthogonal matrices.

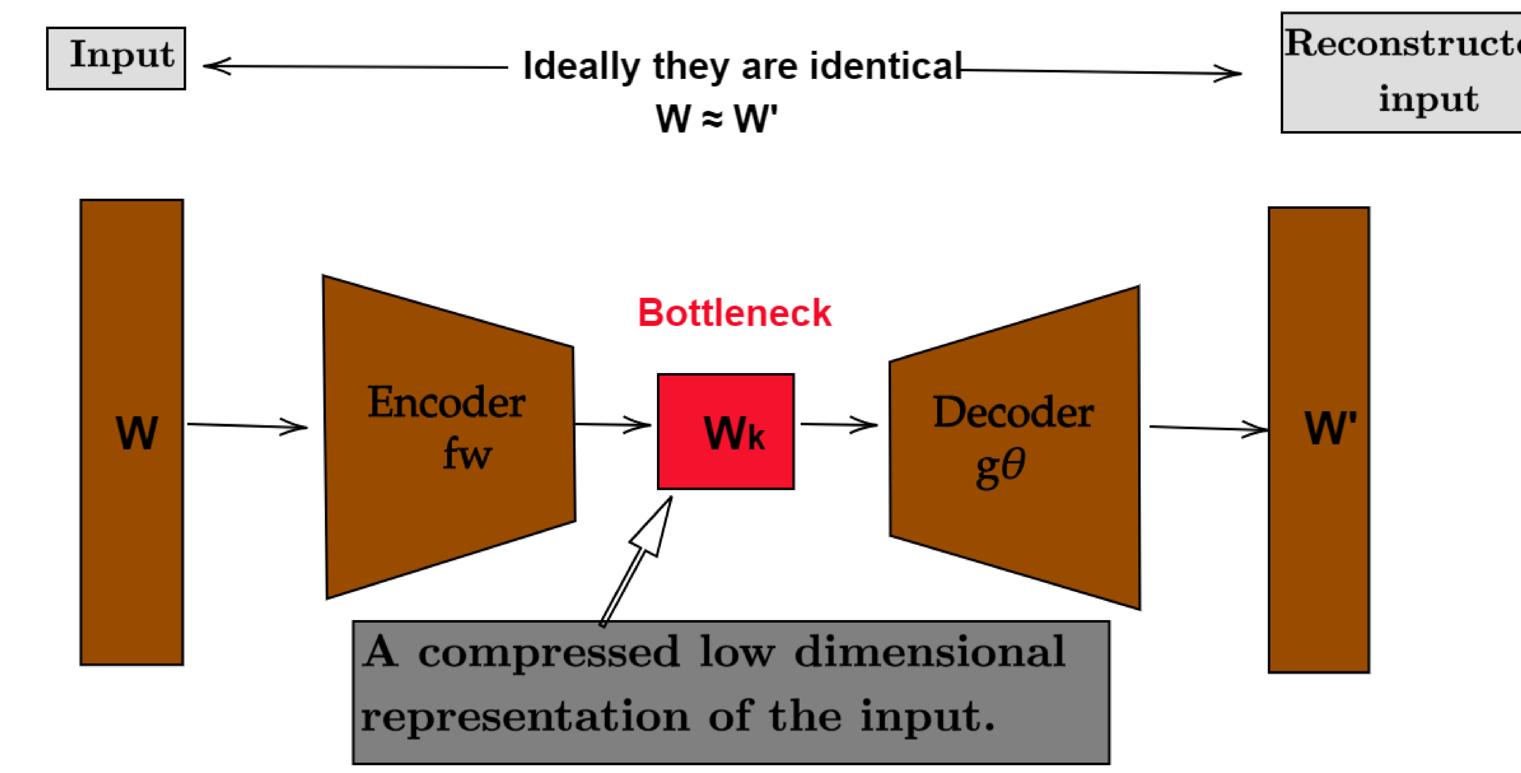


To reduce the dimensionality of the given matrix W_0, we truncate it to retain only the k most important singular values, as they represent the amount of information captured by each singular vector in U and V.



2. Autoencoders

Autoencoder is an artificial neural network-based model that learns efficient representations of data by capturing the most important features while ignoring noise and irrelevant information. An autoencoder has the following parts:



1. **Encoder:** It is a part of the network that compresses the input into a lower-dimensional latent space while extracting the essential features of the input data;
2. **Bottleneck:** It is the lower-dimensional hidden layer where the important features of the input data are captured and represented.
3. **Decoder:** It reconstructs the input data from the compressed representation in the lower-dimensional space to produce an output that is approximately similar to the input data.

By decoding the information encoded in the lower-dimensional space, the decoder attempts to capture the essential features of the input while minimizing the loss of information.

REFERENCES

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al.
The falcon series of open language models.
arXiv preprint arXiv:2311.16867, 2023.

[2] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al.
Mistral 7b.
arXiv preprint arXiv:2310.06825, 2023.

[3] Samruddhi Kahu and Reena Rahate.
Image compression using singular value decomposition.
International Journal of Advancements in Research & Technology, 2(8):244–248, 2013.

[4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al.
Llama 2: Open foundation and fine-tuned chat models.
arXiv preprint arXiv:2307.09288, 2023.

[5] Yasi Wang, Hongxun Yao, and Sicheng Zhao.
Auto-encoder based dimensionality reduction.
Neurocomputing, 184:232–242, 2016.

[6] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang.
A survey on model compression for large language models.
arXiv preprint arXiv:2308.07633, 2023.

nlmakenne@ulige.be