Exploring Low-Rank and Sparse Approximations for Transformers

Ashwin De Silva, **Esther Whang**

JOHNS HOPKINS UNIVERSITY

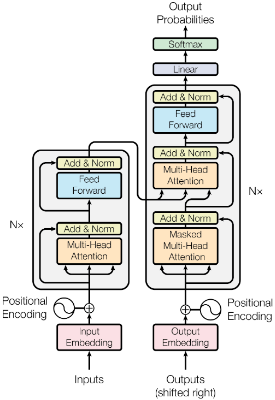
Introduction to Transformers

Transformers are the workhorse of modern deep learning models.

Have major applications in computer vision and natural language processing.

Able to model long range dependencies between different parts of an input sequence.

Downside: high computational costs during inference/finetuning



Compression through Low-Rank Approximation

How to **Reduce Computational Costs** of Transformers while **Maintaining their Performance**?

Assume Transformer is Low-Ranked!

How Valid is this Low-Rank Assumption?

Low-Rank Decomposition: LoRA

Pretrained models have low Intrinsic Dimensionality[1]

Basis for idea that change in weights during model adaptation also has a low "intrinsic rank"

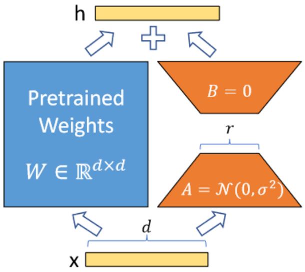


Figure: LoRA: Low-Rank Adaptation of Large Language Models[2]

[1] Aghajanyan et al., *Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning*.
[2] Hu et al., *LoRA: Low-Rank Adaptation of Large Language Models*.

Low-Rank Decomposition: SVD Variations

Standard SVD on weight matrices results in worse performance

Approach: Weigh parameters based on impact on the task performance

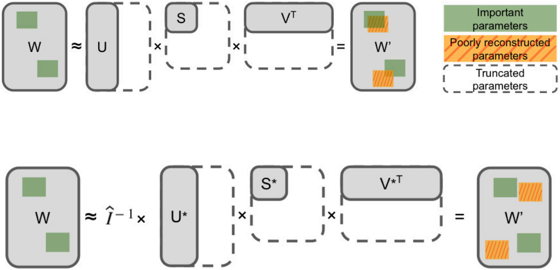


Figure: Fisher-Weighted SVD[3]

Aim for functional equivalence, not just structural similarity

[3] Hsu et al., *Language model compression with weighted low-rank factorization*.

Low-Rank Decomposition: Low-Rank just for MHA

A recent approach applying different compression methods to different modules

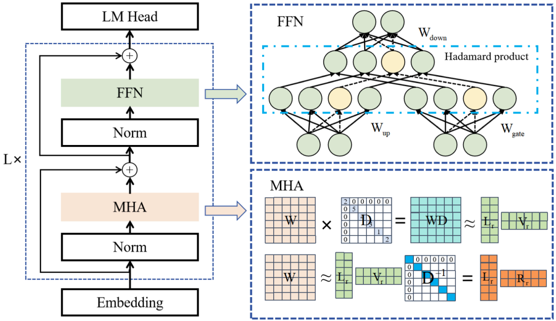


Figure: LoRAP: Low-Ranked matrix approximation And structured Pruning[4]

[4] Li et al., *LoRAP: Transformer Sub-Layers Deserve Differentiated Structured Compression for Large Language Models*.

Project Goals

- Investigate the spectrums of trained attention matrices
- Explore low-rank + sparse approximations for attention
- Compare original and compressed versions

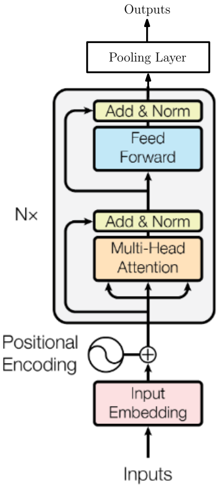
RoBERTa

Uses the same architecture as BERT (Bidirectional Encoder Representations from Transformers)

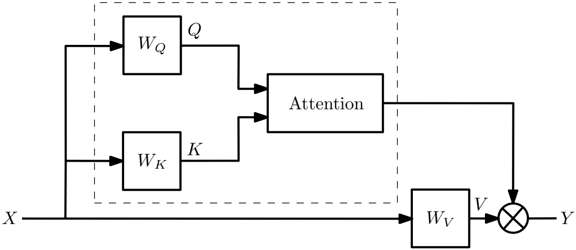
Attention layer parameters
$d = 768, h = 12, N = 12$

Pre-trained using masked language model (MLM) and next sentence prediction (NSP) objectives

For downstream tasks (e.g. sentiment classification), plug a trainable MLP in place of the pooling layer



Attention Operation

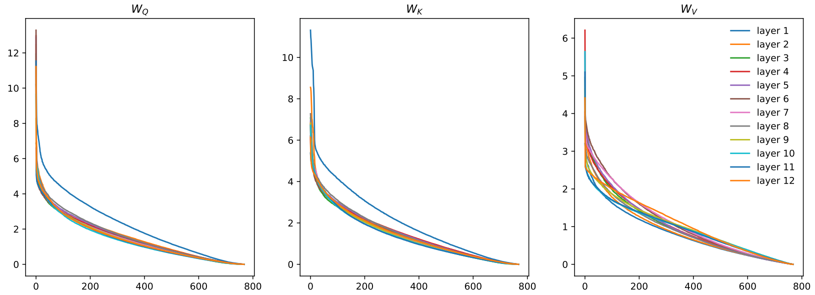


$$\text{Attention}(Q, K, V) = \underbrace{\text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)}_{\text{Attention Probs.}} V$$

$X \in \mathbb{R}^{n \times d},\ W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$

$n =$ sequence length
$d =$ model dimension

Initial observations and intuition



Pre-trained W_K, W_Q, W_V matrices are not strictly low-rank!

Initial observations and intuition

Since $Q = XW_Q, K = XW_K$, Attention operation can be re-written as,

$$\text{Attention}(X) = \text{Softmax}\left(\frac{XBX^\top}{\sqrt{d}}\right)V,$$

where $B = W_Q W_K^\top$

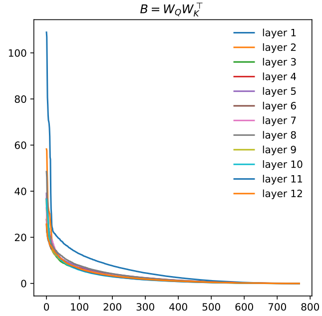
Initial observations and intuition

Since $Q = XW_Q, K = XW_K$, Attention operation can be re-written as,

$$\text{Attention}(X) = \text{Softmax}\left(\frac{XBX^\top}{\sqrt{d}}\right)V,$$

where $B = W_Q W_K^\top$

B has a lower rank compared to the original attention matrices.



Truncated SVD

Low-rank approximation:

$$\min_{U,V} \|W - UV\|_F^2$$
$$\text{s.t. } \text{rank}(U) = \text{rank}(V) \leq r$$

Can be solved using truncated SVD:

$$\text{SVD}(W) = L\Sigma R$$

Retain the r-largest singular values

$$\hat{W} = (L_r \Sigma_r) R_r = UV$$

Robust PCA

Low-rank approximation with sparse noise:

$$\min_{L,S} \text{rank}(L) + \lambda \|S\|_0$$
$$\text{s.t. } W = L + S$$

Can be solved using Principle Component Pursuit[5] (Robust PCA) via ALM:

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1$$
$$\text{s.t. } W = L + S$$

[5] Candès et al., "Robust principal component analysis?"

GreGoDec

Low-rank approximation with sparse and dense noise:

$$\min_{U,V,S} \text{rank}(UV) + \lambda \|S\|_0$$
$$\text{s.t. } \|W - UV - S\|_F^2 \leq \epsilon$$

Can be solved using greedy bilateral smoothing[6]:

$$\min_{U,V,S} \|W - UV - S\|_F^2 + \lambda \|S\|_1$$
$$\text{s.t. } \text{rank}(U) = \text{rank}(V) \leq r$$

Has successful applications in compressing CNNs[7]

[6] Zhou et al., "Greedy bilateral sketch, completion & smoothing"; Zhou et al., "Godec: Randomized low-rank & sparse matrix decomposition in noisy case".

[7] Yu et al., "On compressing deep models by low rank and sparse decomposition".

Single-head attention experiments

Architecture:

- Modified randomly initialized RoBERTa with $d = 768, h = 1, N = 6$ followed by a MLP with 1 hidden layer as the classifier

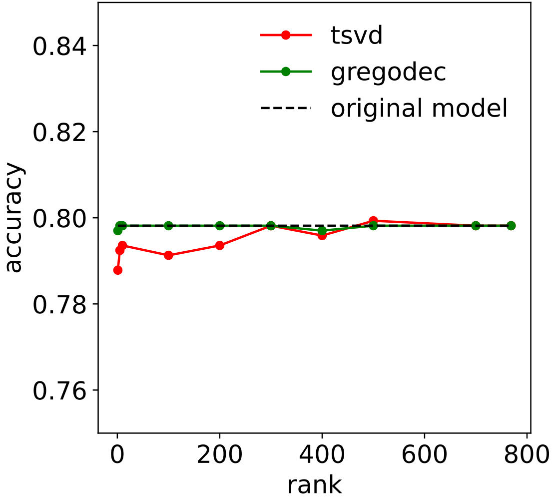
Dataset:

- SST-2 from the GLUE benchmark (sentiment classification)
- Train size: 67.3k, Test size: 872

Method:

- Train the model on SST-2 from scratch.
- Compute and approximate/compress B
- Compare with the original model and different ranks.

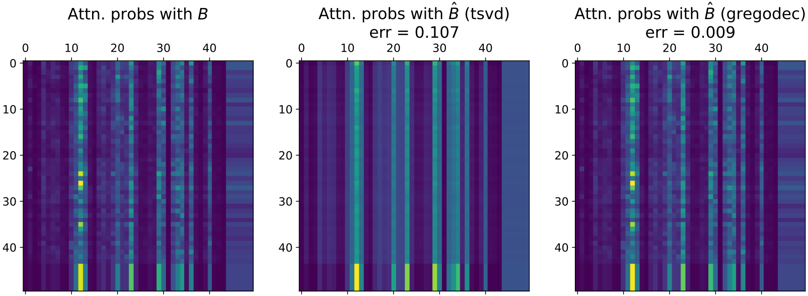
Results



tsvd
gregodec
original model
0.84
0.82
0.80
0.78
0.76
accuracy
0
200
400
600
800
rank

Results

attention scores for “*there’s enough melodrama in this magnolia primavera to make pta proud yet director muccino’s characters are less worthy of puccini than they are of daytime television*”



Multi-head attention experiments

Architecture:

- Original pre-trained RoBERTa with $d = 768, h = 12, N = 12$

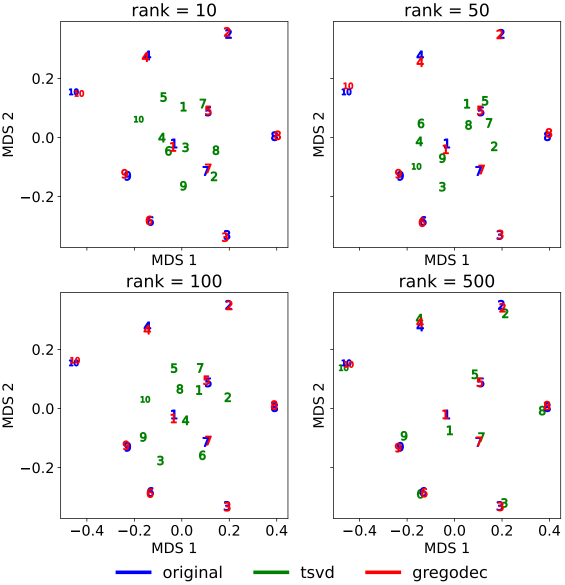
Dataset:

- SST-2 from the GLUE benchmark (sentiment classification)

Method:

- Approximate W_Q, W_K, W_V
- Compare with the original model and different ranks.

Results


rank = 10
rank = 50
rank = 100
rank = 500
MDS 1
MDS 2
0.2
0.0
−0.2
−0.4
0.4
original
tsvd
gregodec

Future Directions

Promoting attention probability preservation

$$\min_{U,V,S} \frac{1}{2m} \left\| \text{softmax}(\tilde{X} B \tilde{X}^\top) - \text{softmax}(\tilde{X}(UV+S)\tilde{X}^\top) \right\|_F^2 + \lambda \|S\|_1$$

$$\text{s.t. } \frac{1}{2}\|B - UV - S\|_F^2 \le \epsilon, \ \text{rank}(UV) \le r$$

Experiments with LoRA

- Combining the pretrained compressed transformer with the LoRA update
- Could lead to more efficient finetuning

Extend single-head results to multi-head

- Will require a non-trivial extension (matrix B vs tensor B)
- Higher-order SVD

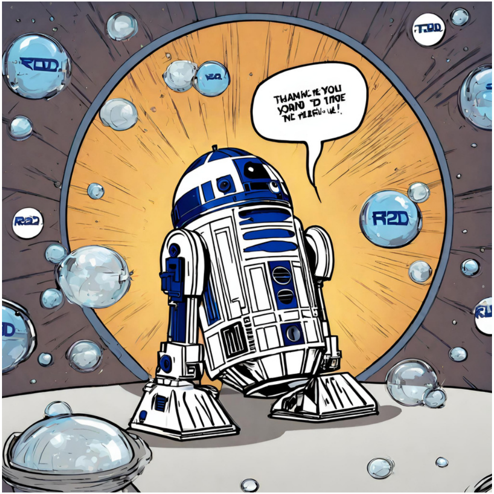
Going beyond RoBERTa

- Large language models

Code

Code for our implementations and experiments are available at:
`https://github.com/Laknath1996/cs-project`

Thank you!



Appendix

Algorithm 1: Greedy Bilateral (GreB) Paradigm

Input: Object function f; rank step size Δr; power K;
tolerance τ; observations of data matrix X

Output: low-rank matrix UV (and sparse S)

Initialize $V \in \mathbb{R}^{r_0 \times n}$ (and S);

while *residual error* $\leq \tau$ **do**

  for $k \leftarrow 1$ **to** K **do**

    Greedy Bilateral Sketch: sequentially compute (9);

    Greedy Bilateral Completion: sequentially compute (13);

    Greedy Bilateral Smoothing: sequentially compute (17);

  end

  Calculate the top Δr right singular vectors v (or Δr-dimensional random projections) of $\partial f / \partial V$ (given in (10), (14) and (18) for different problems);

  Set $V := [V; v]$;

end

Appendix

$$\begin{cases} U_k = (X - S_{k-1}) V_{k-1}^T \left(V_{k-1} V_{k-1}^T \right)^\dagger, \\ V_k = \left(U_k^T U_k \right)^\dagger U_k^T (X - S_{k-1}), \\ S_k = \mathcal{S}_\lambda (X - U_k V_k), \end{cases} \tag{15}$$

$$\frac{\partial \|X - UV - S\|_F^2}{\partial V} = X - UV - S. \tag{18}$$

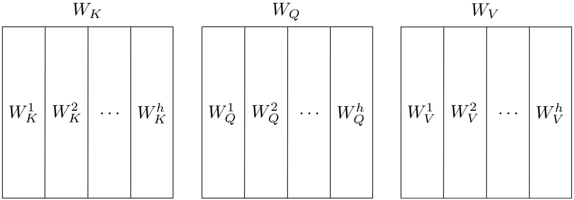
Compression ratio

Let $r = \text{rank}(UV)$ and $s = \|S\|_0$

$$\begin{aligned}\text{compression ratio} &= \frac{\text{\# parameters to represent } \hat{B}}{\text{\# parameters to represent } W_Q, W_K} \\ &= \frac{2 \times (d \times r) + s}{2 \times (d \times d)}\end{aligned}$$

When $s = 0$ and $r = 1$, compression ratio is $1/d$

Attention Operation



$$\begin{aligned} \text{MHA}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W_O \\ \text{head}_i &= \text{Attention}(XW_Q^i, XW_K^i, XW_V^i) \end{aligned}$$

$$W_Q^i, W_K^i, W_V^i \in \mathbb{R}^{d \times d/h},\ W_O \in \mathbb{R}_{d \times d}$$

Results

attention scores for "*may be far from the best of the series, but it's assured, wonderfully respectful of its past and thrilling enough to make it abundantly clear that this movie phenomenon has once again reinvented itself for a new generation.*"