Efficient Weight Matrix Compression

via Kronecker Product Decomposition

Project Report

January 25, 2026

Abstract

This report details the methodology for approximating a large weight matrix W as the Kronecker product of two smaller matrices, A and B. By solving the minimization problem $\min ||W - A \otimes B||_F$, we achieve significant parameter reduction. We utilize the **Pitsianis Rearrangement** method, which transforms the non-linear approximation problem into a Rank-1 Singular Value Decomposition (SVD) problem. A complete step-by-step numerical example is provided to illustrate the workflow.

1 The Core Concept

1.1 The Objective

In modern neural networks, weight matrices (W) can be excessively large, leading to high storage costs and slow inference speeds. Our goal is to "compress" a large matrix W of size $m \times n$ into two smaller factors, A ($m_1 \times n_1$) and B ($m_2 \times n_2$), such that:

$$W \approx A \otimes B \tag{1}$$

Where $\otimes$ denotes the Kronecker Product.

1.2 The Kronecker Product Definition

The Kronecker product creates a block-structured matrix. If A is a 2×2 matrix, $A \otimes B$ is defined as:

$$A \otimes B = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{bmatrix} \tag{2}$$

This implies that if W is a perfect Kronecker product, it is composed of sub-blocks, where every sub-block is simply a scaled version of the matrix B.

2 Methodology: The Algorithm Workflow

To find the optimal matrices A and B, we follow the method proposed by Van Loan & Pitsianis (1993).

2.1 Step 1: Partitioning (Gridding)

We conceptually divide the large matrix W into an $m_1 \times n_1$ grid of blocks. Each block is of size $m_2 \times n_2$.

2.2 Step 2: The Pitsianis Rearrangement ($\mathcal{R}$)

This is the most critical step. We transform the matrix W into a new matrix $\tilde{W}$ (read as "W-tilde").

1. We take each sub-block of W.
2. We **vectorize** the block (flatten it into a single column/row vector).
3. We stack these vectors to form the rows of $\tilde{W}$.

Mathematical Insight

The minimization of the error $||W - A \otimes B||_F$ is mathematically equivalent to finding the best **Rank-1 Approximation** of the rearranged matrix $\tilde{W}$.

2.3 Step 3: SVD Decomposition

We compute the Singular Value Decomposition (SVD) of $\tilde{W}$:

$$\tilde{W} \approx \sigma \cdot u \cdot v^T$$

- The **Right Singular Vector** (v) represents the elements of matrix B (the pattern).
- The **Left Singular Vector** (u) represents the elements of matrix A (the scaling factors).
- The **Singular Value** (σ) represents the magnitude/energy.

2.4 Step 4: Reshaping

We take the vectors u and v and reshape them back into the dimensions of A and B.

3 Detailed Numerical Example

Let us apply this workflow to a specific 4×4 weight matrix to find factors A (2×2) and B (2×2).

3.1 Input Matrix

$$W = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 3 & 4 & 6 & 8 \\ 5 & 10 & 1 & 2 \\ 15 & 20 & 3 & 4 \end{bmatrix} \tag{3}$$

3.2 Step A: Partitioning

We treat W as a 2×2 grid of blocks.

$$W = \begin{bmatrix} \mathbf{W_{11}} & \mathbf{W_{12}} \\ \mathbf{W_{21}} & \mathbf{W_{22}} \end{bmatrix}$$

The blocks are identified as:

$$W_{11} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \text{(Top-Left)}, W_{21} = \begin{bmatrix} 5 & 10 \\ 15 & 20 \end{bmatrix} \quad \text{(Bottom-Left)}$$

$$W_{12} = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}, \quad \text{(Top-Right)}, W_{22} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad \text{(Bottom-Right)}$$

3.3 Step B: Rearrangement ($\tilde{W}$)

We vectorize each block into a row and stack them. *(Note: Following the standard convention for this example, we proceed by block-columns: $W_{11}, W_{21}, W_{12}, W_{22}$).*

- Row 1 (from W_{11}): $[1, 2, 3, 4]$
- Row 2 (from W_{21}): $[5, 10, 15, 20]$
- Row 3 (from W_{12}): $[2, 4, 6, 8]$
- Row 4 (from W_{22}): $[1, 2, 3, 4]$

The rearranged matrix is:

$$\tilde{W} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 10 & 15 & 20 \\ 2 & 4 & 6 & 8 \\ 1 & 2 & 3 & 4 \end{bmatrix} \tag{4}$$

3.4 Step C: SVD Solution

We observe the pattern in $\tilde{W}$:

- Row 1 is the base: $[1, 2, 3, 4]$
- Row 2 is $5\times$ Row 1.

- Row 3 is $2\times$ Row 1.
- Row 4 is $1\times$ Row 1.

The SVD extracts the principal vectors:

- **Pattern Vector (v)**: $[1,2,3,4]^T$ (Corresponds to B)
- **Scaling Vector (u)**: $[1,5,2,1]^T$ (Corresponds to A)

3.5 Step D: Reshaping to Final Matrices

1. Reconstructing Matrix A: We take the scaling vector $u = [1,5,2,1]^T$. We fill the matrix column-by-column (matching the order we stacked the blocks).

$$A = \begin{bmatrix} 1 & 2 \\ 5 & 1 \end{bmatrix} \tag{5}$$

2. Reconstructing Matrix B: We take the pattern vector $v = [1,2,3,4]^T$.

$$B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \tag{6}$$

4 Verification

To verify the accuracy of our decomposition, we compute $A \otimes B$:

$$A \otimes B = \begin{bmatrix} 1 \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} & 2 \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \\ \\ 5 \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} & 1 \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 3 & 4 & 6 & 8 \\ 5 & 10 & 1 & 2 \\ 15 & 20 & 3 & 4 \end{bmatrix}$$

This matches the original matrix W exactly. The approximation error is zero.

5 Conclusion

This report demonstrated the mathematical workflow for decomposing a weight matrix using Kronecker products. By utilizing the Pitsianis rearrangement and SVD, we successfully extracted the underlying factors A and B, reducing the parameter count from 16 to 8 (a 50% compression).