DOCS: Quantifying Weight Similarity for Deeper Insights into Large Language Models

Zeping Min & Xinshang Wang

Analyzing Weight Matrices via DOCS

Large Language Models (LLMs) based on transformer architectures have achieved remarkable successes, yet interpreting their internal mechanisms remains challenging. Traditional similarity analysis techniques primarily focus on representations, which do not directly reflect the similarity between weight matrices due to two key issues:

- **Representation vs. Weights:** Residual connections, defined by $\mathbf{y} = \mathcal{F}(\mathbf{x}, \mathcal{W}) + \mathbf{x}$, inherently create correlated layer representations, even when underlying weight matrices differ significantly. Thus, similar layer representations do not imply similar weight structures.
- **Limitations with Orthogonal Matrices:** Existing methods like Canonical Correlation Analysis (CCA), SVCCA, and linear CKA fail to discriminate effectively between orthogonal matrices, yielding identical similarity scores regardless of true structural differences. Since orthogonal matrices frequently appear in trained LLMs, these methods fall short of accurately characterizing weight similarity.

To overcome these limitations, we introduce the *Distribution of Cosine Similarity (DOCS)*, a novel matrix-similarity index designed specifically for weight matrices. DOCS computes pairwise cosine similarities between corresponding vectors within weight matrices and analyzes their distribution, effectively capturing structural differences even among orthogonal matrices. Experiments demonstrate DOCS provides clearer insights into the similarities and differences of LLM weight matrices compared to existing methods.

Distribution of Cosine Similarity

Our DOCS method quantifies the similarity between neural network components by analyzing the alignment of their weight matrices. Each component (such as feed-forward layers or attention heads) is represented by a matrix whose columns correspond to parameter vectors (e.g., neuron weights, attention patterns). DOCS computes the similarity through the following procedure:

- **Cosine Similarity Computation:** We first calculate the cosine similarity between all parameter vector pairs from two matrices X and Y, yielding a similarity matrix C, where each entry C_{jk} represents the similarity between vectors X_j and Y_k.
- **Identify Maximum Alignments:** For each parameter vector, we record its highest absolute cosine similarity with vectors in the other matrix.
- **Model with Gumbel Distribution:** We model these maximum similarities with Gumbel distributions, estimating location parameters u_X and u_Y via maximum likelihood.
- **Calculate DOCS Similarity Index:** Averaging these two location parameters gives the scalar DOCS similarity index S_{DOCS}, a measure between 0 and 1 reflecting overall component similarity.

Unlike methods relying on global matrix correlations (e.g., Canonical Correlation Analysis, SVCCA, linear CKA), DOCS emphasizes strong individual vector alignments. This approach effectively captures functional similarities and differences between neural network components, enhancing structural analysis in deep learning models.

Mathematical Properties

Method	PT Invariance	Symmetry	IS Invariance	Reflexivity	Behavior on Orthogonal Matrices
Linear Regression	✓	✗	✓	✓	Constant
CCA (R^2_{CCA})	✓	✓	✓	✗	Constant
CCA ($\bar{\rho}_{\text{CCA}}$)	✓	✓	✓	✓	Constant
SVCCA (R^2_{SVCCA})	✓	✓	✓	✓	Constant (assuming $T_X = T_Y = I$)
SVCCA ($\bar{\rho}_{\text{SVCCA}}$)	✓	✓	✓	✓	Constant (assuming $T_X = T_Y = I$)
Linear HSIC	✓	✗	✗	✗	Dimension-Dependent
Linear CKA	✓	✓	✓	✓	Constant
DOCS (Ours)	✓	✓	✓	✓	**Discriminative**

Experimental Results