# Embedding Matrix Evolution in Vision Transformers

John Drotos
Corey Toler-Franklin, Graphics Imaging and Light Measurement Lab, Barnard College, Columbia University

## Abstract

This study investigates the behavior of embedding matrices in Vision Transformers when applied to medical image datasets. By comparing a pretrained transformer model to the same model retrained on a cancer detection dataset, we aim to understand how transformers adapt to the complex and subtle features of medical images. We extracted the query, key, and value matrices from both models and compared their similarities through multiple metrics. The results revealed significant global and local patterns, shedding light on the nature of embedding matrices. These findings lay the groundwork for future improvements in transformer-based medical image analysis.

## Introduction

- Medical images pose a significant challenge for machine learning models because they are often feature-poor and lack strong color or texture contrast (Fig. 1). These characteristics make it difficult for machine learning models to effectively detect tumors, particularly when the tumors are small or indistinct. [1]
- Vision Transformers have demonstrated success in image recognition due to their ability to model complex, non-local dependencies. [2] However, the intermediate stages of transformers are often left as a black box, so understanding how these stages learn and represent patterns could pave the way for breakthroughs in medical object detection.

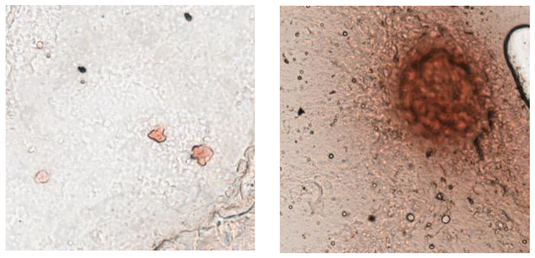


Figure 1: Two images from the lab's dataset, both containing tumors. These images demonstrate the wide scale and texture of tumors, which makes tumor detection challenging for machine learning models.

## Math and Transformer Architecture

Transformer models are comprised of two main mechanisms

1. Multi-Layer Perceptron (MLP): The input matrix X undergoes two linear transformations, $W_{inp}$ and $W_{out}$, interspersed by a non-linear activation function GELU.

$$\mathrm{MLP}(\mathbf{X}) = \mathrm{GELU}(\mathbf{X}\mathbf{W}_{\mathrm{inp}})\mathbf{W}_{\mathrm{out}}$$

2. Multi-Head Self-Attention (MHSA): The input matrix X is passed through several attention heads, $A^1 \ldots A^f$, which compute the attention scores for each token. The attention scores are derived from the query (Q) and key (K) matrices, representing the relationships between tokens. These scores are then used to weight the value (V) vectors, $W_V$, producing the final output. Figs. 2 and 3 demonstrate this process.

$$\mathrm{MHSA}(\mathbf{X}) = \mathrm{hconcat}\left[\mathbf{A}^1\mathbf{X}\mathbf{W}^1_{V_{\mathrm{attn}}}, \ldots, \mathbf{A}^f\mathbf{X}\mathbf{W}^f_{V_{\mathrm{attn}}},\right]\mathbf{W}_{\mathrm{out}}$$

The transformer architecture alternates between MHSA and MLP layers, feeding the output of MHSA into the MLP to refine and process the token embeddings, as shown in Fig. 4.

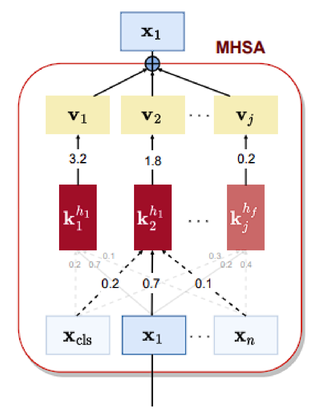



Figure 2: A vector-level diagram of Multi-Head Self-Attention, from "Analyzing Vision Transformers for Image Classification in Class Embedding Space" (https://arxiv.org/abs/2310.18969)

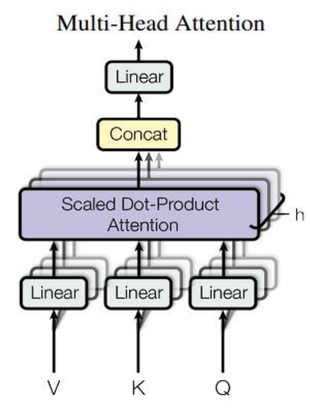



Figure 3: A matrix-level diagram of Multi-Head Attention, from "Attention is All You Need" (https://arxiv.org/abs/1706.03762)

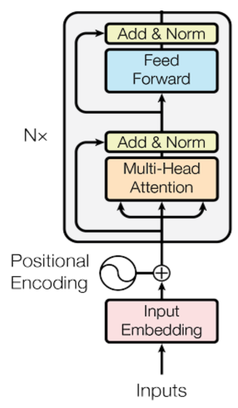
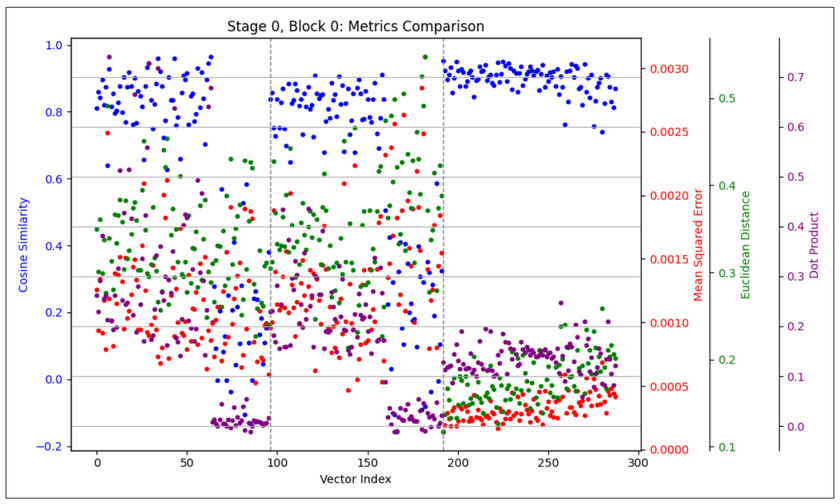
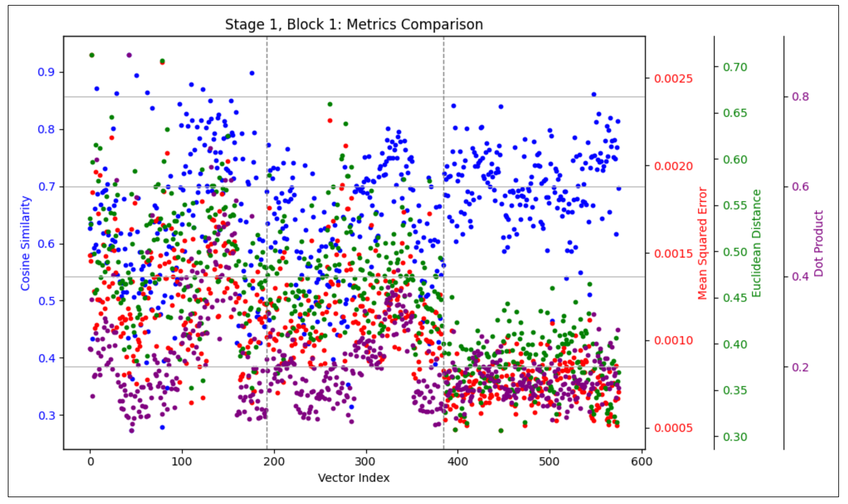
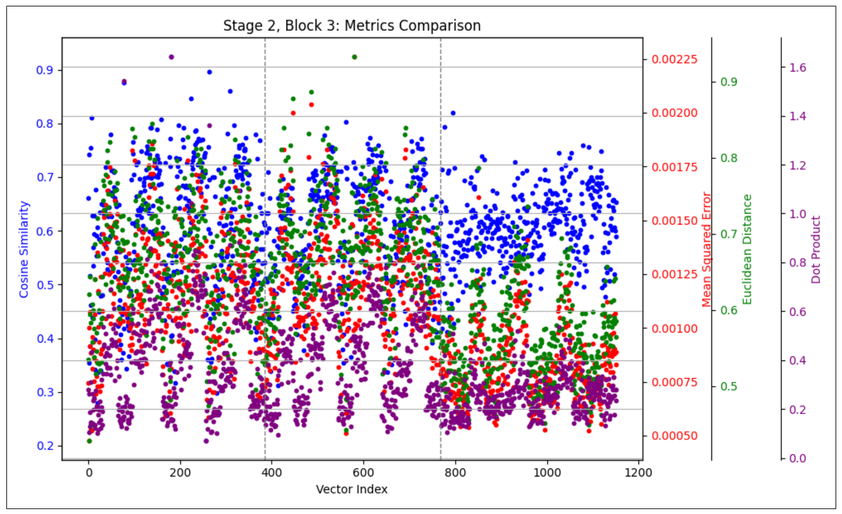



Figure 4: The basic architecture of a Transformer, showing the relationship between the MHSA and MLP (Feed Forward) mechanisms, from "Attention is All You Need" (https://arxiv.org/abs/1706.03762)



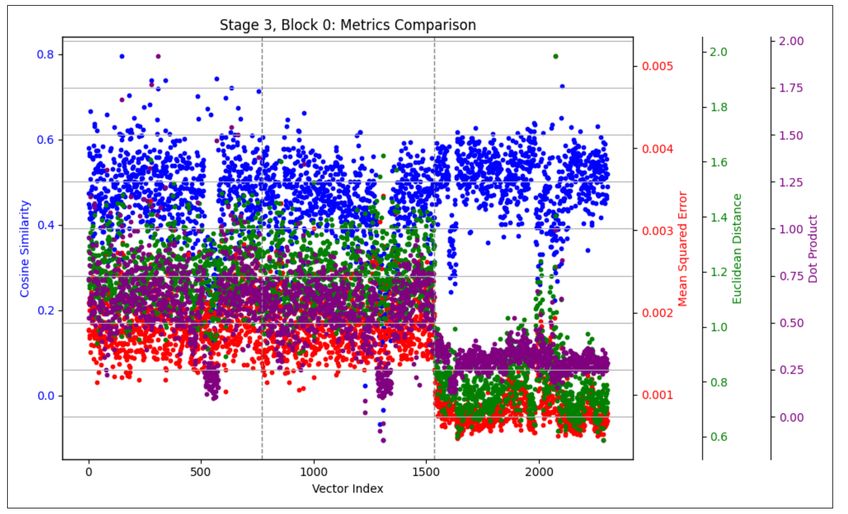




Figure 5: Vector comparisons between the embedding matrices at several points throughout the transformer

## Methods

### Model Selection and Training

The control model was a Mask-RCNN with a SWIN transformer backbone, pretrained on the ImageNet-1k dataset using the MMDetection framework. The same architecture was initialized with these weights for the test model and trained on the lab's medical dataset for 6 epochs in a Docker container.

### Extraction of Embedding Matrices

After training, embedding matrices from all 24 blocks (across 4 stages) of both models were extracted. To compare them, four similarity metrics were computed for each corresponding vector pair: cosine similarity (1 minus cosine), Euclidean distance, mean squared error, and dot product (Figure 5).

## Analysis

### Global Patterns

- Both MSE and Euclidean distance dropped significantly in the value vectors across all stages and blocks.
- The dot product also decreased in the value matrix, while cosine similarity showed minimal global change.

### Local Patterns

- Clustering and significant changes were observed within sub-matrices, especially in Stage 2.
- Notable patterns include a drop in cosine similarity at the end of the key and query sub-matrices in Stage 0 Block 0, and acute drops in dot product and cosine similarity in Stage 3 Block 0.

## Discussion

- Cosine similarity, Euclidean distance, and MSE measure vector similarity inversely, while dot product increases with vector alignment.
- Lower values in the value sub-matrix suggest greater similarity, but lower dot products indicate a possible contradiction.
- Dot product correlates with vector magnitude, suggesting that the value vectors might have less magnitude than the key and query vectors.
- Local clusters highlight distinct patterns within each sub-matrix, but their cause and significance remain unclear.
- These findings suggest potential directions for future research into vector magnitude and local pattern causes.

## Conclusion

This study reveals significant evolution in the embedding matrices of a Vision Transformer when trained on a medical image dataset. Future work could test whether similar patterns are observed across other medical datasets and could explore ways to interpret local clustering effects. Ultimately, this research sets the stage for designing more specialized AI models capable of improving medical image analysis.

### Sources

1. Salgia, Ravi, et al. "Quantifying cancer: More than just a numbers game." *Trends in Cancer* 7.4 (2021): 267-269
2. Prince, Simon J. D. *Understanding Deep Learning*. The MIT Press, 2023.
3. Vilas, Martina G., Timothy Schaumlöffel, and Gemma Roig. "Analyzing Vision Transformers for image classification in class embedding space." *Advances in neural information processing systems* 36 (2024).
4. Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).