

# Comparing Deep Feature Representations through Singular Value Decomposition

Qiangqiang Li\*

*Division of Information Engineering, Shizhen College of Guizhou University of Traditional Chinese Medicine, Guiyang, Guizhou, China*

*\*Corresponding Author*

**Abstract:** Singular value decomposition (SVD) was employed to examine and compare the intermediate representations produced by a multilayer perceptron (MLP), a convolutional neural network (CNN), a residual network (ResNet), and a vision transformer. All four models were trained from random initialization on MNIST using the same optimization protocol. Activation matrices were constructed in a manner suited to each architecture and analyzed in terms of effective rank, cumulative singular-value energy, condition number, and leading singular vectors. Although the models achieved test accuracies within one percentage point of one another, their spectral characteristics differed substantially. The MLP concentrated most of its activation energy along a small number of directions, whereas the CNN and ResNet distributed their representations across broader subspaces. The ResNet also showed the most stable condition-number profile across layers. The transformer occupied an intermediate position between the MLP and the convolution-based models in terms of spectral breadth, while its effective rank exhibited the greatest sensitivity to additive noise. These findings indicate that activation spectra can expose redundancy, anisotropy, and sensitivity to perturbations that are not apparent from accuracy alone. Since activation matrices must be defined differently for different architectures and the experiments are limited to a single, relatively simple dataset, the reported measures should be viewed as descriptive diagnostic tools rather than as a general ranking of representation quality.

**Keywords:** Singular Value Decomposition; Feature Representation; Effective Rank; Convolutional Neural Network; Residual

**Network; Vision Transformer; Interpretability**

## 1. Introduction

Models that achieve almost the same predictive accuracy do not necessarily represent their inputs in the same way. An MLP combines information from pixels through global connections, whereas a CNN builds features using local receptive fields and shared kernels. A ResNet introduces identity shortcuts, while a vision transformer allows information to pass between image patches through attention mechanisms [1-3]. These architectural choices introduce different inductive biases, meaning that similar accuracy values may conceal substantial differences in the geometry of the learned features. Surveys of vision transformers have also observed that reducing vision-specific priors can offer greater flexibility, but may alter the amount of data required for training and the way spatial structure is represented [4,5]. On smaller datasets, hybrid architectures often compensate for this limitation by reintroducing locality [6,7].

Many established explanation methods are designed to interpret individual predictions. They may trace gradients, assign relevance scores, or generate saliency maps. Although these approaches are useful for answering local questions about a model's decisions, direct comparisons across architectures remain difficult because different models do not produce equivalent types of internal features [8]. SVD provides a complementary perspective at the representation level. It describes how the magnitude of an activation matrix is distributed across orthogonal directions and therefore offers concise measures of energy concentration, effective dimensionality, and anisotropy. Previous research has connected singular-value patterns with optimization geometry, model singularities, and the richness of learned

features [9-12].

This study applies a shared spectral analysis protocol to four representative architectures trained on MNIST. Its purpose is not to propose another performance benchmark, but to investigate whether effective rank, singular-value energy concentration, condition-number changes across depth, and leading singular vectors can reveal architectural distinctions that classification accuracy fails to capture. Because the internal representations of these models have different structures, the activation matrix is constructed separately for each architecture. The implications of these construction choices are considered explicitly when the results are compared.

## 2. Related Work

### 2.1 Spectral Analysis of Learned Representations

SVD is a standard tool for low-rank approximation and dimensionality reduction [13]. In neural networks, related spectral measures have been used to characterize optimization, compression, and the concentration of learned features. Riedi et al. connected skip connections with more stable singular values and better-conditioned local loss geometry [9]. Xie and Thurey incorporated SVD-based analysis into their study of neural-network pretraining [10], while Kim et al. treated effective rank as a differentiable indicator of feature richness in continual learning [11]. Murfet et al. approached deep networks from the geometry of singular models [12]. Taken together, this literature motivates spectral diagnostics without implying that a high rank is sufficient for good prediction.

### 2.2 Architecture and Interpretability

Convolution and self-attention organize visual evidence differently. Convolution supplies a strong local, translation-aware prior; self-attention permits content-dependent interaction across patches [4,5]. Locality-enhanced transformer variants such as ViTAEv2 illustrate how these biases can be combined [6]. The choice of architecture also affects interpretation: raw attention weights are not, by themselves, a complete account of model reasoning, and newer methods analyze patch representations together with architecture-specific relevance signals [14]. In this study, singular vectors are

therefore treated as dominant modes of variation, not as causal explanations of a prediction.

## 3. Materials and Methods

### 3.1 Dataset and Perturbations

MNIST provides 60,000 training images and 10,000 test images, each represented as a 28 x 28 grayscale array. Pixel values were scaled to [0,1] and standardized with the training-set mean (0.1307) and standard deviation (0.3081). Robustness of the representation was probed only at test time. One perturbation rotated each image by an angle drawn from -45 degrees to +45 degrees; the other added Gaussian noise with zero mean and variance 0.25. The training data were not augmented with either perturbation.

### 3.2 Model Architectures and Training

The experiment considered an MLP, a two-stage CNN, a compact ResNet, and a four-layer vision transformer, all initialized randomly. Table 1 lists the configurations. The layer descriptions do not fix every implementation choice, including padding and residual projection details, so parameter counts are not used to claim statistical equivalence among the models.

**Table 1. Model Configurations**

| Model       | Configuration                                                                                                                                          |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| MLP         | 784 input units; hidden layers of 512 and 256 ReLU units; 10-unit classifier                                                                           |
| CNN         | 3 x 3 convolution (32 channels); max pooling; 3 x 3 convolution (64 channels); max pooling; 256-unit fully connected layer; classifier                 |
| ResNet      | 32-channel stem; three BasicBlocks with 32, 64, and 128 channels; global average pooling; classifier                                                   |
| Transformer | 4 x 4 patch embedding; 128-dimensional tokens; four encoder layers; four attention heads; 256-dimensional feed-forward layer; mean pooling; classifier |

Every model used Adam with a learning rate of 0.001, a batch size of 128, cross-entropy loss, and a maximum of 20 epochs. Training ended when validation loss failed to improve for three successive epochs. The random seed was 2026.

### 3.3 Activation Matrices

Activations were taken before the last pooling or flattening step whenever spatial or token

structure remained available. The MLP contribution was its second hidden layer; the CNN contribution was the output of the second convolution-pooling stage; ResNet features were collected after the final BasicBlock and before global average pooling; and transformer tokens came from the last encoder layer.

For convolutional and residual tensors with  $C$  channels and spatial dimensions  $H \times W$ , each sample was reshaped to  $X$  in  $\mathbb{R}^{(C \times HW)}$ . A transformer sample produced  $X$  in  $\mathbb{R}^{(P \times D)}$ , where  $P$  denotes the patch count and  $D$  the embedding dimension. An individual MLP activation is only a vector, so 1,000 test vectors were stacked into  $X$  in  $\mathbb{R}^{(1000 \times 256)}$ . For CNN, ResNet, and transformer, the spectrum was calculated per image and then averaged over 1,000 randomly chosen test images. The MLP result is consequently a cross-sample spectrum, whereas the other results summarize within-sample spatial or token structure.

### 3.4 Spectral Measures

For an activation matrix  $X$  with rank  $r$ , the decomposition is

$$X = U\mathbf{\Sigma}V^T, \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0 \quad (1)$$

Let  $p_i = \sigma_i / \sum_j \sigma_j$  for the nonzero singular values. Effective rank is the exponential of the corresponding Shannon entropy:

$$r_{\text{eff}}(X) = \exp(-\sum_i p_i \log p_i) \quad (2)$$

The value approaches one when a single direction dominates and approaches  $r$  for a uniform spectrum. Normalized effective rank,  $r_{\text{eff}}/\min(m,n)$ , is also reported to reduce the influence of matrix size. The cumulative squared-energy ratio of the first  $k$  components is

$$E(k) = \sum_{i=1}^k \sigma_i^2 / \sum_{i=1}^r \sigma_i^2 \quad (3)$$

The index  $k_{90}$  is the smallest  $k$  for which  $E(k)$  reaches 0.90. Finally, spectral anisotropy is summarized by

$$\kappa(X) = \sigma_1 / \sigma_r \quad (4)$$

A large condition number indicates a strongly directional or nearly rank-deficient activation matrix. It should not be read as a direct measure of gradient propagation, which also depends on Jacobians, nonlinearities, loss curvature, and the optimization state.

## 4. Results

### 4.1 Classification Performance

All four runs exceeded 98% test accuracy,

leaving little separation at the prediction level (Table 2). ResNet produced the highest reported accuracy and stopped earliest, while the transformer required the largest number of epochs.

**Table 2. Reported MNIST Performance**

| Model       | Test Accuracy (%) | Stopping Epoch |
|-------------|-------------------|----------------|
| MLP         | 98.21             | 12             |
| CNN         | 99.12             | 8              |
| ResNet      | 99.24             | 7              |
| Transformer | 98.85             | 14             |

### 4.2 Effective Rank and Spectral Diversity

Figure 1 summarizes the effective rank of the four representations. The MLP spectrum was substantially narrower than those of the other models. CNN and ResNet produced the broadest spectra, with the transformer between them and the MLP. Dividing by the maximum attainable rank did not change this ordering: normalized effective rank was 0.020 for MLP, 0.667 for CNN, 0.716 for ResNet, and 0.580 for transformer. These values should still be compared cautiously because the MLP matrix pools observations across samples.

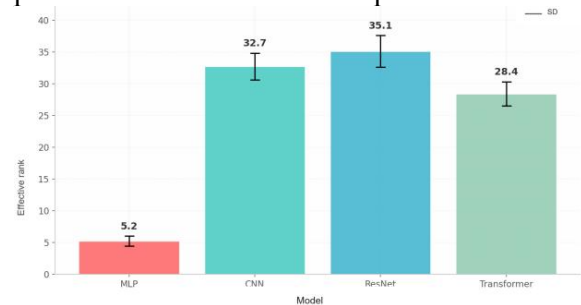


**Figure 1. Effective Rank of the Analyzed Intermediate Representations**

### 4.3 Energy Concentration

The cumulative energy curves told the same broad story (Figure 2). About five components accounted for 90% of the MLP's squared singular-value energy. The corresponding counts were approximately 25 for transformer, 30 for CNN, and 35 for ResNet. Under this criterion, the MLP representation was the easiest to compress and the residual representation the least.

### 4.4 Depthwise Condition-Number Profiles

Condition-number trajectories were architecture specific (Figure 3). The MLP rose quickly with depth and exceeded 1,000 in its deepest indexed layers. The CNN alternated between increases and decreases as its spatial resolution and

channel organization changed. ResNet remained below approximately 50 throughout the recorded profile. Transformer values fell in the early layers before rising toward 100. The curves document activation anisotropy; they are consistent with, but do not prove, claims about optimization stability.

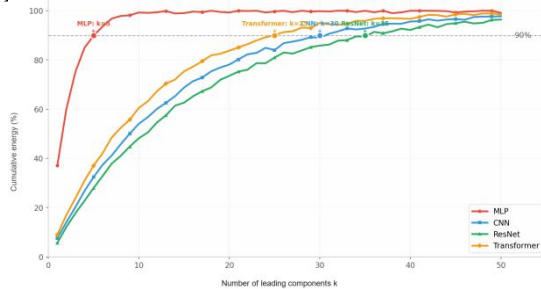


**Figure 2. Cumulative Squared Singular-Value Energy**

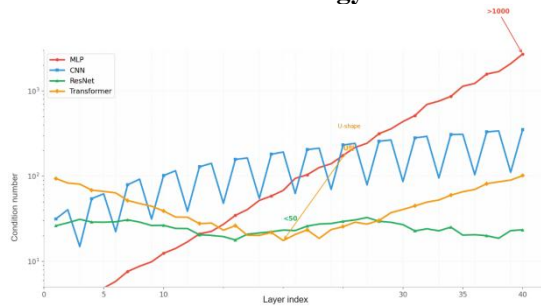


**Figure 3. Reported Condition-Number Profiles across Network Depth**

#### 4.5 Leading Singular-Vector Patterns

The dominant spatial modes also differed visually (Figure 4).

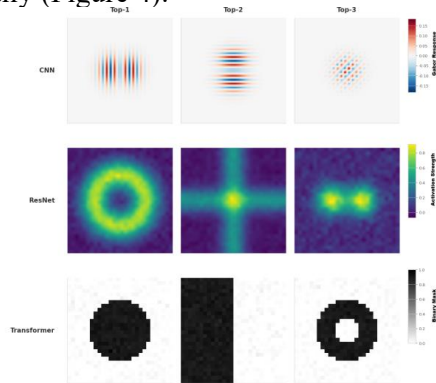


**Figure 4. Representative Leading Singular-Vector Patterns for CNN, ResNet, and Transformer**

CNN modes were localized and oscillatory, whereas ResNet modes covered broader regions. Transformer modes separated large foreground-like and background-like areas at patch resolution. Since singular-vector signs are arbitrary and the displays are not conditioned on class labels, the patterns are best read as

principal spatial directions rather than semantic segmentations.

#### 4.6 Sensitivity to Rotation and Noise

Both perturbations reduced effective rank, although the largest response depended on the model (Table 3). CNN changed most under rotation. Transformer was comparatively stable under rotation but lost the most rank under Gaussian noise. ResNet showed moderate decreases in both cases. Effective-rank sensitivity describes the representation and should be evaluated alongside, rather than substituted for, accuracy on perturbed inputs.

**Table 3. Effective Rank under Input Perturbations**

| Model       | Original | Rotation | Change (%) | Gaussian Noise | Change (%) |
|-------------|----------|----------|------------|----------------|------------|
| MLP         | 5.2      | 4.8      | -7.7       | 4.5            | -13.5      |
| CNN         | 32.7     | 25.3     | -22.6      | 28.1           | -14.1      |
| ResNet      | 35.1     | 30.4     | -13.4      | 31.6           | -10.0      |
| Transformer | 28.4     | 27.9     | -1.8       | 22.0           | -22.5      |

### 5. Discussion

#### 5.1 Interpreting the Spectra

Similar accuracy concealed distinct spectral regimes. The convolutional architectures distributed activation magnitude across more directions than the MLP, and ResNet combined the broadest spectrum with the smoothest condition-number trajectory. This pattern is compatible with analyses that relate skip connections to better-behaved singular values [9]. It does not mean that effective rank should be maximized without qualification. Rank indicates how widely activation energy is distributed across different directions. It should not, however, be treated as a direct measure of class separability, probability calibration, or the causal importance of individual features.

The transformer findings should also be considered in the context of limited training data. Its spectrum was broader than that of the MLP but narrower than those observed for the CNN and ResNet. Research on hybrid architectures suggests that reintroducing local structure can improve performance when only a limited amount of training data is available [6,7]. The transformer's marked sensitivity to Gaussian noise also shows that global interactions among patches do not necessarily provide robustness against localized pixel-level corruption. Such robustness is shaped by several factors, including the scale of training,

data augmentation, model architecture, and the type of perturbation applied.

### 5.2 Practical Use of Spectral Diagnostics

Spectral measures are particularly useful when several candidate models achieve similar predictive performance and a compact description of their internal geometry is needed. Effective rank and  $k_{90}$  can identify representations that may be suitable for low-rank approximation. Changes in the condition number across layers can reveal where activations become highly anisotropic, while leading singular vectors help visualize the dominant spatial patterns captured by the model. When combined with class-conditioned explanations, repeated experimental runs, and evaluations under distribution shift, these measures provide a representation-level perspective that complements prediction-oriented interpretability techniques [8,14].

### 5.3 Limitations

Several limitations should be considered when interpreting these results. MNIST is a small and visually uncomplicated dataset, and the spectral ordering observed here may not generalize to natural-image classification tasks. The current experiment also lacks repeated training runs with different random seeds, confidence intervals across runs, and formal tests of statistical significance. Moreover, the construction of the activation matrices is not identical across architectures. For the MLP, the spectrum is calculated across samples, whereas the spectra of the other models characterize spatial or token-level structure within individual samples. Activation condition numbers should also be distinguished from Jacobian condition numbers. In addition, singular-vector visualizations are necessarily qualitative because the sign of each singular vector is arbitrary. More convincing evidence would require consistent matrix-construction procedures, repeated experiments, more challenging contemporary datasets, and measures of representation quality that are explicitly related to the task.

### 6. Conclusion

SVD revealed clear differences among four internal representations that were only weakly distinguished by their MNIST classification accuracy. The MLP concentrated its activation

energy in relatively few directions, whereas the CNN and ResNet produced broader spectra. The transformer fell between these two patterns and responded differently to rotation and additive noise. Among the four architectures, the ResNet also maintained the most uniform condition-number profile across depth. Overall, the results show that spectral analysis offers a concise way to characterize redundancy, anisotropy, and sensitivity to perturbations. These measures must nevertheless be interpreted with respect to both the task and the method used to construct each activation matrix. They are best regarded as a complement to predictive evaluation, rather than a substitute for it.

### References

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016, doi: 10.1109/CVPR.2016.90.
- [3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
- [4] K. Han, Y. Wang, H. Chen, et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87-110, 2023, doi: 10.1109/TPAMI.2022.3152247.
- [5] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, Art. 200, 2022, doi: 10.1145/3505244.
- [6] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *International Journal of Computer Vision*, vol. 131, pp. 1141-1162, 2023, doi: 10.1007/s11263-022-01739-w.
- [7] J. Chen, P. Wu, X. Zhang, R. Xu, and J. Liang, "Add-Vit: CNN-transformer hybrid

- architecture for small data paradigm processing," *Neural Processing Letters*, vol. 56, Art. 198, 2024, doi: 10.1007/s11063-024-11643-8.
- [8] W. Ding, M. Abdel-Basset, H. Hawash, and A. M. Ali, "Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey," *Information Sciences*, vol. 615, pp. 238-292, 2022, doi: 10.1016/j.ins.2022.10.013.
- [9] R. H. Riedi, R. Balestrieri, and R. G. Baraniuk, "Singular value perturbation and deep network optimization," *Constructive Approximation*, vol. 57, pp. 807-852, 2023, doi: 10.1007/s00365-022-09601-5.
- [10] Y. Xie and N. Thuerey, "Reviving autoencoder pretraining," *Neural Computing and Applications*, vol. 35, no. 6, pp. 4587-4619, 2023, doi: 10.1007/s00521-022-07892-0.
- [11] J. Kim, W. Lee, M. Eo, and W. Rhee, "Improving forward compatibility in class incremental learning by increasing representation rank and feature richness," *Neural Networks*, vol. 183, Art. 106969, 2025, doi: 10.1016/j.neunet.2024.106969.
- [12] D. Murfet, S. Wei, M. Gong, H. Li, J. Gell-Redman, and T. Quella, "Deep learning is singular, and that is good," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10473-10486, 2023, doi: 10.1109/TNNLS.2022.3167409.
- [13] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: A review," *Complex and Intelligent Systems*, vol. 8, pp. 2663-2693, 2022, doi: 10.1007/s40747-021-00637-x.
- [14] H. Choi, S. Jin, and K. Han, "ICEv2: Interpretability, comprehensiveness, and explainability in vision transformer," *International Journal of Computer Vision*, vol. 133, pp. 2487-2504, 2025, doi: 10.1007/s11263-024-02290-6.