Analysis of Feature Extraction Power of VGG16 Before and After Transfer Learning on MNIST Dataset

Shahriar Hasan
Id: 2010776115
Department of Computer Science and Engineering
University of Rajshahi

June 21, 2025

1. Introduction

Convolutional Neural Networks (CNNs) are powerful deep learning models for image recognition. VGG16 is a popular CNN architecture pre-trained on ImageNet, capable of learning rich feature representations. In this study, we explore the feature extraction power of VGG16 before and after fine-tuning on the MNIST handwritten digit dataset. High-dimensional feature vectors extracted from VGG16 are visualized using three dimensionality reduction techniques: PCA, t-SNE, and LDA.

2. Methodology

2.1 Dataset and Preprocessing

We used the MNIST dataset consisting of 60,000 training and 10,000 test grayscale images of handwritten digits (0–9). Since VGG16 expects RGB images of size 32x32, we:

- Converted grayscale images to RGB by replicating channels.
- Resized images from 28x28 to 32x32.
- Normalized pixel values to [0, 1].

2.2 VGG16 Model and Transfer Learning

We used the pre-trained VGG16 model with ImageNet weights and removed the top classification layers. The model was extended by:

- Adding a global average pooling layer.
- Adding two dense layers (512 and 256 units with ReLU).
- Using Dropout regularization (0.5 and 0.3).

- Output layer with 10 softmax units for MNIST classification.

Two models were used:

- **Before Transfer Learning**: Extracted features from frozen VGG16.
- **After Transfer Learning**: Fine-tuned the custom top layers on MNIST.

2.3 Feature Extraction

We extracted high-dimensional feature vectors:

- From `block5_pool` (before training).
- From the global average pooling layer (after training).

2.4 Dimensionality Reduction

To visualize the extracted features, we applied:

- **PCA (Principal Component Analysis)** — linear projection.
- **t-SNE (t-distributed Stochastic Neighbor Embedding)** — non-linear manifold learning.
- **LDA (Linear Discriminant Analysis)** — supervised projection maximizing class separability.

3. Training Performance

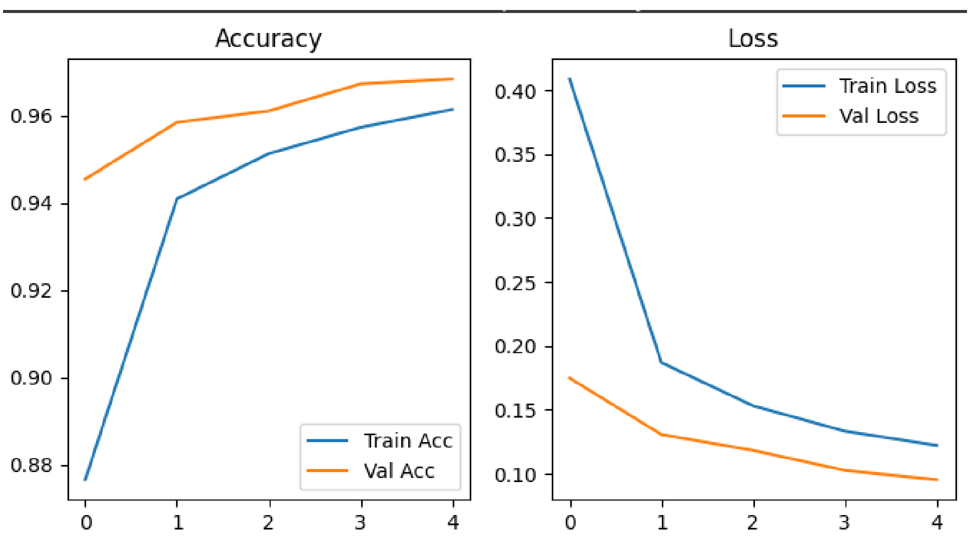

Figure 1: Training and Validation Accuracy and Loss over 5 Epochs

The model quickly converges, achieving high accuracy on MNIST. This demonstrates the power of transfer learning even with frozen base layers.

4. Feature Visualizations

We visualized 1,000 randomly selected test samples using PCA, t-SNE, and LDA.

4.1 Before Transfer Learning

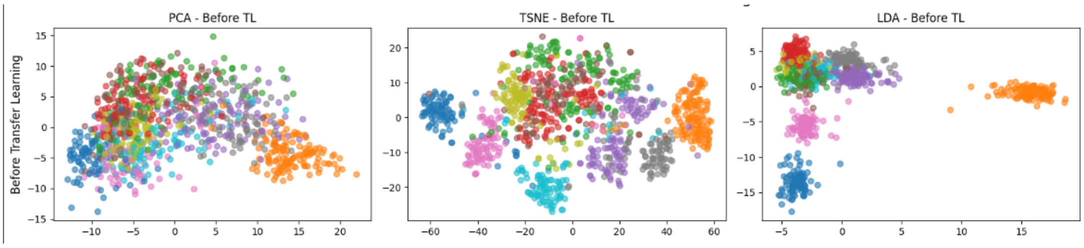

Figure 2: 2D Feature Projections Before Transfer Learning (PCA, t-SNE, LDA)

4.2 After Transfer Learning

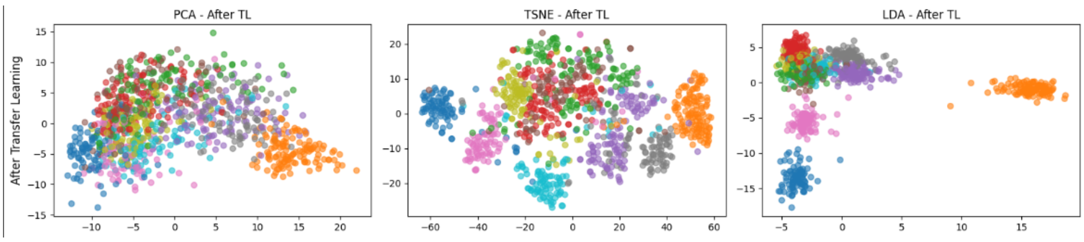

Figure 3: 2D Feature Projections After Transfer Learning (PCA, t-SNE, LDA)

5. Analysis

- Before training, features extracted from the VGG16 backbone show poor separation between classes, especially for PCA and LDA.
- After training on MNIST, features become more discriminative.
- t-SNE and LDA demonstrate significant improvements in clustering digits after transfer learning.
- PCA shows increased variance explained by the first few components after transfer learning.

6. Conclusion

VGG16 pretrained on ImageNet provides a strong feature extractor. However, transfer learning on domain-specific data like MNIST significantly improves class-wise separation in feature space. Visualizing feature vectors using PCA, t-SNE, and LDA confirms that fine-tuning enhances feature discriminability.