# Analysis of Feature Extraction Power of MobileNetV2 Before and After Transfer Learning

Student ID: 2010676102

June 21, 2025

## 1 Introduction

This report details an experiment to investigate the feature extraction capabilities of a pre-trained Convolutional Neural Network (CNN). Specifically, we examine MobileNetV2, pre-trained on the ImageNet dataset, and its ability to extract meaningful features for a distinct task: MNIST handwritten digit recognition. The core of the analysis involves visualizing high-dimensional feature vectors in a 2D plane using dimensionality reduction techniques, both before and after fine-tuning the model on the MNIST dataset. This approach allows for a qualitative assessment of how transfer learning adapts the learned feature representations.

## 2 Experimental Methodology

The experiment was conducted using Python with TensorFlow/Keras, and the code is provided here - [Link]

### 2.1 Model

The chosen pre-trained CNN was MobileNetV2, utilizing its weights learned from the ImageNet dataset. The top classification layer of MobileNetV2 was excluded.

### 2.2 Dataset

The MNIST dataset, consisting of grayscale images of handwritten digits (0-9), was used as the target dataset.

### 2.3 Process

The experimental process involved the following key steps:

1. **Dataset Preparation:** MNIST images (originally 28x28 grayscale) were preprocessed by resizing them to 48x48 pixels and converting them to 3-channel (RGB) format to be compatible with

MobileNetV2's input shape. Pixel values were normalized to the [0,1] range. A subset of 2000 test images was used for feature extraction and visualization.

2. **Feature Extraction (Before Transfer Learning):** Features were extracted from the output of the `GlobalAveragePooling2D` layer, which was appended to the base MobileNetV2 model (with ImageNet weights). The preprocessed MNIST test images were fed through this feature extractor.

3. **Dimensionality Reduction and Visualization (Before TL):** The extracted high-dimensional feature vectors were reduced to 2D using three techniques: Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Locally Linear Embedding (LLE). The 2D representations were plotted, with points colored according to their true MNIST digit labels.

4. **Transfer Learning (Fine-tuning):**
   - The MobileNetV2 base model had its original top layers removed.
   - A new classification head was added, consisting of a `GlobalAveragePooling2D` layer, followed by a `Dense` layer with softmax activation for 10 classes (MNIST digits).
   - The base MobileNetV2 layers were initially frozen, except for the last 10 layers, which were unfrozen for fine-tuning.
   - The model was trained on the preprocessed MNIST training set for 5 epochs using the Adam optimizer and categorical cross-entropy loss.

5. **Feature Extraction (After Transfer Learning):** Features were extracted from the analogous `GlobalAveragePooling2D` layer of the fine-tuned model using the same MNIST test images.

6. **Dimensionality Reduction and Visualization (After TL):** The process from step 3 was repeated for the features extracted from the fine-tuned model.

# 3 Results and Discussion

## 3.1 Feature Visualization Before Transfer Learning

The dimensionality reduction techniques were applied to the features extracted from the original MobileNetV2 (pre-trained on ImageNet) when fed with MNIST digit images. The resulting visualizations are shown in Figure 1.

- **PCA:** The PCA plot (Figure 1(1.a)) shows very poor separation of the digit classes. While the digit '1' forms a somewhat distinct group on the right, the remaining digits are heavily intermingled. This indicates that the principal components of variance in the feature space do not align with the different digit classes.
- **t-SNE:** The t-SNE plot (Figure 1(1.b)) reveals some local structure, but the different digit classes are still largely intermingled with no clear global separation. The lack of distinct clustering indicates that the ImageNet-derived features, prior to fine-tuning, do not naturally group MNIST digits into their respective classes.
- **LLE:** The LLE plot (Figure 1(1.c)) also demonstrates poor separation. It projects the data onto a manifold that reflects the local linear relationships but fails to separate the digit classes into meaningful clusters.

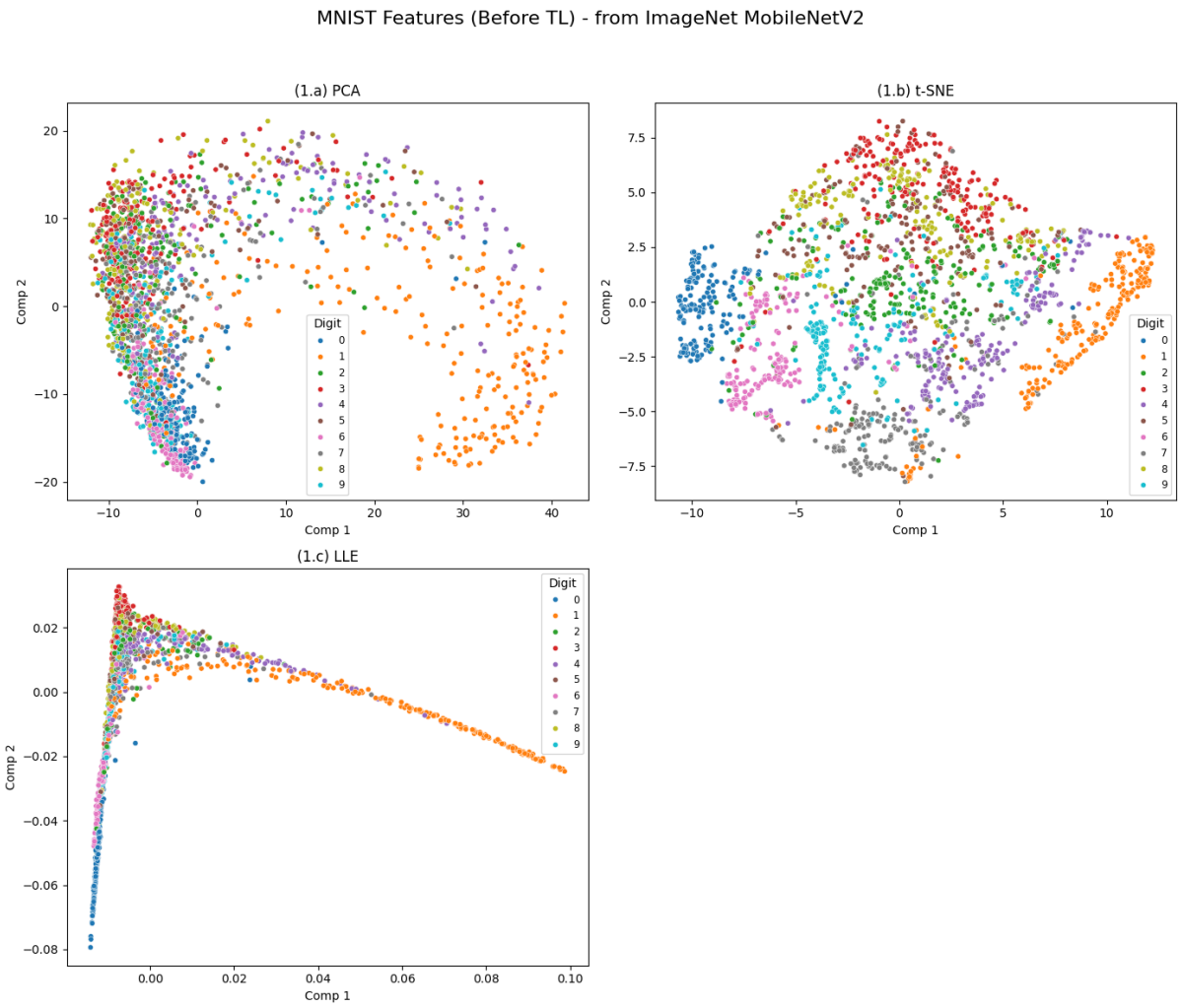


Figure 1: 2D visualization of MNIST features extracted from the pre-trained MobileNetV2 model (before fine-tuning) using PCA, t-SNE, and LLE.

The overall poor separation before transfer learning demonstrates that the feature extraction power of the ImageNet-trained MobileNetV2, while strong for general visual tasks, is not directly transferable or optimal for the specific nuances of MNIST digit classification without adaptation.

### 3.2 Feature Visualization After Transfer Learning (Fine-tuning)

After fine-tuning MobileNetV2 on the MNIST dataset for 5 epochs, features were again extracted and visualized, as shown in Figure 2.

- **PCA:** As seen in Figure 2(2.a), PCA now shows a significantly improved separation of digit classes compared to the pre-tuning scenario. While some overlap remains, distinct regions for each digit begin to emerge, indicating that fine-tuning has pushed the feature representations of different digits further apart in the high-dimensional space.
- **t-SNE:** The t-SNE plot generated after fine-tuning (Figure 2(2.b)) exhibits a dramatic improvement. Clear, distinct clusters corresponding to each of the 10 MNIST digits (0-9) are now visible.

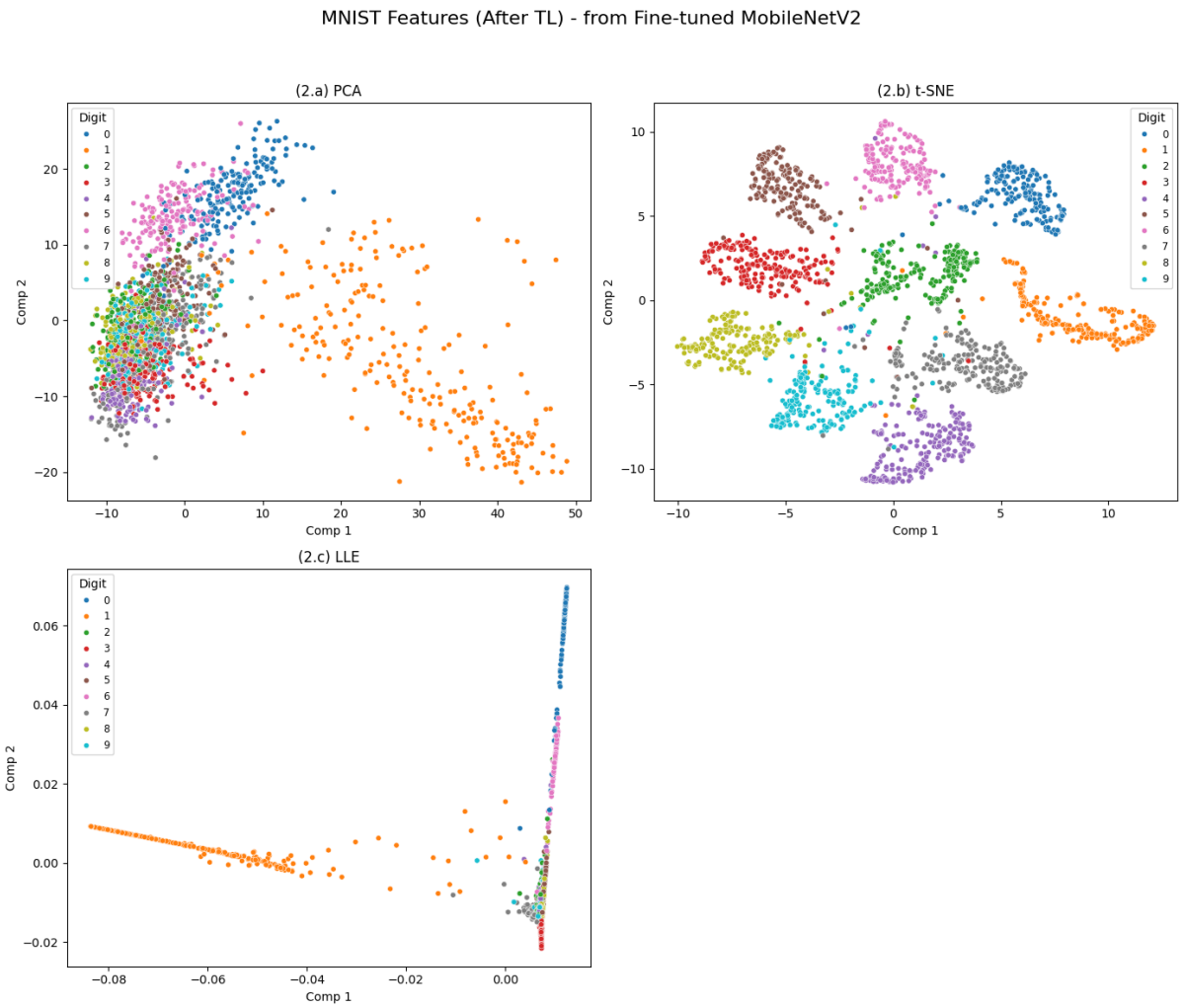


Figure 2: 2D visualization of MNIST features extracted from the fine-tuned MobileNetV2 model using PCA, t-SNE, and LLE.

This indicates that the transfer learning process successfully adapted the network's feature extraction layers to learn representations that are highly discriminative for the MNIST task.

- **LLE:** The LLE plot (Figure 2(2.c)) also shows a marked improvement. The points for each digit class are now largely grouped together, forming distinct structures along the learned manifold.

### 3.3 Discussion on Feature Extraction Power

The comparison of feature visualizations before and after transfer learning clearly illustrates the impact of fine-tuning on the model's feature extraction power for the target task.

- **Before Fine-tuning:** As seen in Figure 1, the pre-trained MobileNetV2 possessed a strong general feature extraction capability, but these features were not specific enough to effectively separate MNIST digits. The high-dimensional representations of different digits largely overlapped.
- **After Fine-tuning:** The process of unfreezing some later layers and training on MNIST allowed the model to adapt its feature extractors. The weights in these layers were adjusted to minimize the classification loss on MNIST, thereby learning to produce feature vectors that are separable

for the digit classes. The t-SNE plot in Figure 2 vividly demonstrated this transformation, with the initially mixed feature points reorganizing into well-separated clusters.

- This adaptation highlights that while pre-trained models provide an excellent starting point by capturing general visual hierarchies, their feature extraction power for a new, specific task is significantly enhanced through transfer learning. The network learns to emphasize features relevant to the new task and suppress those that are not.
- In this experiment, t-SNE created significantly more distinct and well-separated clusters for the MNIST digits compared to PCA and LLE, especially after fine-tuning. This is because t-SNE is a non-linear algorithm specifically designed to preserve local neighborhood structures, meaning it focuses on keeping similar data points (like images of the same digit) close together in its 2D map. In contrast, PCA is a linear technique that prioritizes global variance, which doesn't necessarily align with class separability, and LLE focuses on preserving local linear relationships which can be too rigid for the complex variations in handwritten digits. As a result, t-SNE's ability to model complex, non-linear similarities makes it exceptionally effective at visually separating the intricate patterns of the MNIST feature vectors.

The dimensionality reduction techniques, particularly t-SNE, proved effective in visualizing this change in the high-dimensional feature space, making the abstract concept of "feature extraction power" more tangible.

# 4 Conclusion

This experiment successfully demonstrated the feature extraction power of a pre-trained CNN (MobileNetV2) and the transformative effect of transfer learning. Visualizations of feature vectors using dimensionality reduction techniques (PCA, t-SNE, and LLE) showed a marked improvement in class separability after fine-tuning the model on the MNIST dataset. Initially, features extracted by the ImageNet-trained model resulted in poorly separated clusters for MNIST digits. After fine-tuning, the features became highly discriminative, leading to clearly defined clusters for each digit class. This underscores the importance of adapting pre-trained models to new tasks to leverage their learned hierarchical features effectively, thereby significantly enhancing their feature extraction power for the specific domain.