Dimensionality Reduction
Machine Learning Project 2

TZU-CHIEH CHAO

Department of Engineering Education

Master of Science in Applied Data Science

University of Florida

Abstract—This project undertakes a thorough examination of dimensionality reduction methodologies as applied to the CIFAR-10 dataset, assessing their efficacy in both visualization and subsequent classification endeavors. The investigation centers on the evaluation of various strategies aimed at diminishing the high-dimensional image space while maintaining essential discriminative information necessary for classification purposes. Techniques such as Principal Component Analysis (PCA) and manifold learning algorithms were employed to convert the data into lower-dimensional representations, which were subsequently analyzed using a range of classification models. The experimental framework utilized the HiperGator supercomputing infrastructure at the University of Florida, which comprises 32 CPU cores. This setup included baseline models—namely, Support Vector Classifier, Random Forest, and Logistic Regression—operating without dimensionality reduction, which were then juxtaposed with models utilizing features reduced through PCA and manifold learning techniques. This research enhances the understanding of the trade-offs associated with various dimensionality reduction techniques and their implications for image classification tasks, with potential applications in computer vision systems focused on object recognition and scene understanding, where the efficient processing of intricate visual data is of paramount importance.

Keywords—Machine Learning, Images Processing, Principal Component Analysis , Dimension Reduction , Supercomputer

I. INTRODUCTION

Dimensionality reduction is a fundamental aspect of data analysis and machine learning, facilitating the transformation of high-dimensional datasets into lower-dimensional representations while maintaining critical information. This methodology offers several advantages, including decreased computational complexity, diminished noise, and enhanced model performance. The present project investigates a range of dimensionality reduction techniques, notably Principal Component Analysis (PCA) and manifold learning algorithms, in the context of image classification utilizing the CIFAR-10 dataset.

The CIFAR-10 dataset consists of 60,000 color images, each measuring 32x32 pixels, categorized into ten distinct classes, thereby presenting a complex multi-class classification challenge. Through the application of dimensionality reduction, this study seeks to uncover a lower-dimensional feature space that effectively encapsulates the distinguishing characteristics of the images, thereby facilitating efficient and accurate classification. The performance of three widely utilized classifiers—Support Vector Classifier (SVC), Random Forest, and Logistic Regression—will be assessed both with and without the implementation of dimensionality reduction techniques.

In addition to image classification, dimensionality reduction techniques possess considerable potential across various domains, including music streaming. Music streaming services manage extensive datasets characterized by numerous features that describe songs and artists. The application of dimensionality reduction in this context could significantly improve music recommendation systems, genre classification, and music discovery by identifying key features and simplifying data complexity. This study lays the groundwork for further exploration of dimensionality reduction applications in similarly data-intensive environments where efficient and insightful analysis is imperative.

II. METHODOLOGY

A. Dataset Description

The CIFAR-10 dataset contains 60,000 labeled 32×32 RGB images across 10 classes (e.g., airplane, cat, dog), split into 50,000 training and 10,000 test images. It features intra-class variations and inter-class similarities, posing challenges for classification. Commonly used for CNN evaluation, it undergoes normalization and augmentation (flipping, cropping). Models like ResNet-56 achieve ~93-95% accuracy. While its small size enables fast prototyping, low resolution limits real-world applicability. CIFAR-10 remains a key benchmark for computer vision research.

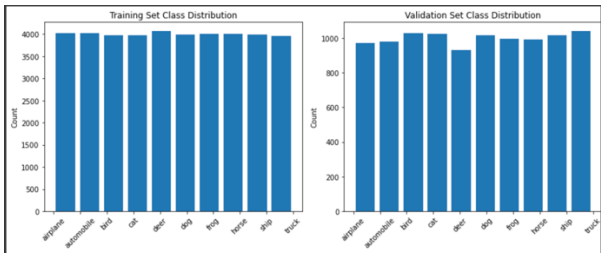
B. Data Preprocessing

The preprocessing pipeline standardizes input images for machine learning models. For datasets like CIFAR-10 (32×32×3 RGB images), the following steps are applied:

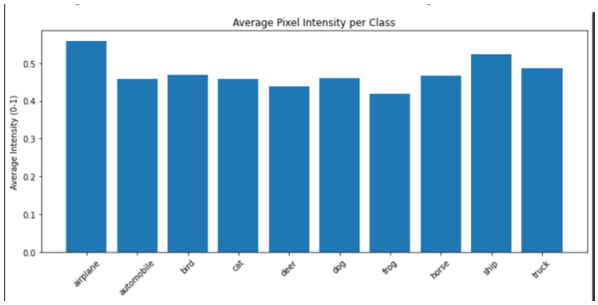
1. **Reshaping**: Flatten each 3D image tensor (32×32×3) into a 2D array (1×3,072 dimensions) using X.reshape(X.shape[0], -1), converting spatial and color channels into a single feature vector per sample. This is essential for models requiring 1D input (e.g., linear classifiers).
2. **Normalization**: Scale pixel values from [0, 255] to [0, 1] by dividing by 255 (X_normalized = X_reshaped / 255.0). This stabilizes training by ensuring uniform numerical ranges across features.

C. *Visualization Analysis*

- Class Distribution - The plot shows that the sample counts for each class are roughly equal and evenly distributed across both the training and validation sets. This indicates that the dataset is relatively balanced concerning class representation, which is beneficial for model training and evaluation as it avoids biases toward specific classes.



- Average Pixel Intensity - The plot reveals variations in average pixel intensity across different classes. For example, "ship" and "airplane" have higher average pixel intensities, possibly due to the presence of bright skies or water bodies in their images. Conversely, "cat" and "dog" exhibit lower average pixel intensities, potentially because of the prevalence of dark fur in their images.



III. Evaluation

A. *Model Development & Evaluation*

Baseline Models (No Dimensionality Reduction)

This section delineates the process of model development, encompassing the selection of classification algorithms, strategies for hyperparameter tuning, and metrics for performance evaluation. We utilized three widely recognized classifiers for image classification: Support Vector Classifier (SVC), Random Forest, and Logistic Regression. These models were subjected to training and evaluation both with and without the application of dimensionality reduction techniques in order to evaluate their performance and efficiency.

Initially, we trained the three classifiers on the original, high-dimensional CIFAR-10 dataset without applying any dimensionality reduction. This established baseline performance for comparison with models trained on reduced feature spaces. Hyperparameter tuning was performed using GridSearchCV with a predefined parameter grid for each classifier. The key hyperparameters considered for each model are as follows:

- **SVC:** kernel (rbf, linear), C (0.1, 1, 10), gamma (scale, auto)
- **Random Forest:** n_estimators (50, 100, 200), max_depth (None, 5, 10)
- **Logistic Regression:** penalty (l2), solver (saga, lbfgs), C (0.01, 0.1, 1, 10), max_iter (100, 200, 500)

The models were evaluated using accuracy and macro-averaged F1-score as performance metrics. Training time was also recorded to assess computational efficiency.

Baseline Models (No Dimensionality Reduction)

	Model	Accuracy	F1 Score	Training Time (s)	Inference Time (s)	Best Parameters
0	SVC	0.4169	0.415975	22.408150	81.441629	{'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
1	Random Forest	0.4692	0.464919	173.004491	0.408138	{'max_depth': None, 'max_features': 'sqrt', 'n...
2	Logistic Regression	0.3337	0.332471	162.928465	0.128320	{'C': 0.01, 'max_iter': 100, 'solver': 'saga'}

SVC : The model had an accuracy of 41.69% and a macro F1-score of 0.4160. Optimal parameters were C=10, gamma='scale', and kernel='rbf'. Training was quicker (22.41 seconds), but inference was slow (81.44 seconds). Like Random Forest, "ship" and "automobile" scored well (F1 0.56 and 0.51), while "bird" and "cat" lagged (0.29 and 0.31). The high inference time makes SVC less practical for real-time applications despite its moderate accuracy.

Random : The model achieved an accuracy of 46.92%, with a macro F1-score of 0.4649. The best parameters were max_depth=None, max_features='sqrt', and n_estimators=200. Training took 173 seconds, and inference was fast at 0.4081 seconds. Performance varied by class: "ship" and "truck" had the highest F1-scores (0.60 and 0.55), while "cat" and "bird" performed poorly (0.33 and 0.35). The model struggled with fine-grained distinctions, suggesting potential improvements with feature engineering or data augmentation.

Logistic : Logistic Regression performed the weakest, with 33.37% accuracy and a 0.3325 F1-score. Best parameters were C=0.01, max_iter=100, and solver='saga'. Training took 162.93 seconds, but inference was efficient (0.1283 seconds). "Ship" and "truck" again led (F1 0.44 and 0.40), while "bird" and "dog" trailed (0.21 and 0.26). The low scores indicate linear models may lack complexity for this task, favoring non-linear alternatives like CNNs.

Dimensionality Reduction with PCA

To reduce the dimensionality of the CIFAR-10 dataset, we employed Principal Component Analysis (PCA). We determined the number of principal components required to explain 90% of the variance in the data. This reduced feature space was then used to train the three classifiers. Hyperparameter tuning was performed similarly to the baseline models, with the addition of

tuning the number of principal components (n_components) for PCA.

SVC(with PCA) :The PCA-reduced SVC model achieved an accuracy of 39.67% with a macro F1-score of 0.3917. The best parameter were pca_n_components=100, svc_C=1, and kernel='rbf'. Training took 39.98 seconds, and inference was moderately fast (2.59 seconds). Performance was uneven: "ship" and "automobile" scored well (F1 0.52 and 0.46), while "bird" and "cat" lagged (0.24 and 0.29). PCA helped reduce dimensionality, but the model still struggled with class imbalances and fine-grained features.

Random Forest(with PCA):This hybrid model had lower accuracy (36.05%) and F1-score (0.3526). Optimal settings were pca_n_components=50, max_depth=20, and n_estimators=100. Training was quick (4.39 seconds), and inference was efficient (0.3358 seconds). "Ship" and "airplane" performed best (F1 0.48 and 0.44), while "bird" and "dog" were weakest (0.22 and 0.27). PCA's dimensionality reduction may have oversimplified features, hurting Random Forest's ability to capture complex patterns.

Logistic Regression (with PCA):The worst performer, with 34.04% accuracy and a 0.3368 F1-score. Best parameters were pca_n_components=50, lr_C=1, and solver='saga'. Training took 9.22 seconds, with fast inference (0.1772 seconds). "Ship" and "truck" led (F1 0.44 and 0.41), while "bird" and "dog" trailed (0.22 and 0.27). PCA's linearity combined with Logistic Regression's simplicity likely limited its ability to handle non-linear class boundaries, making it unsuitable for this task.

Models with PCA Dimensionality Reduction

	Model	Accuracy	F1 Score	Training Time (s)	Inference Time (s)	PCA Components
0	PCA + SVC	0.3967	0.391745	39.982142	2.590395	100
1	PCA + Random Forest	0.3605	0.352634	4.392016	0.335762	50
2	PCA + Logistic Regression	0.3404	0.336801	9.220447	0.177243	50

B. *Manifold Learning*

Manifold Learning for Dimensionality Reduction and Visualization -

To better understand the structure of the CIFAR-10 dataset, we applied two nonlinear manifold learning techniques—**t-SNE** (t-Distributed Stochastic Neighbor Embedding) and **Isomap**—to project the high-dimensional image data into a 2D space for visualization. Both methods were evaluated on a subset of 2,000 samples due to computational constraints.

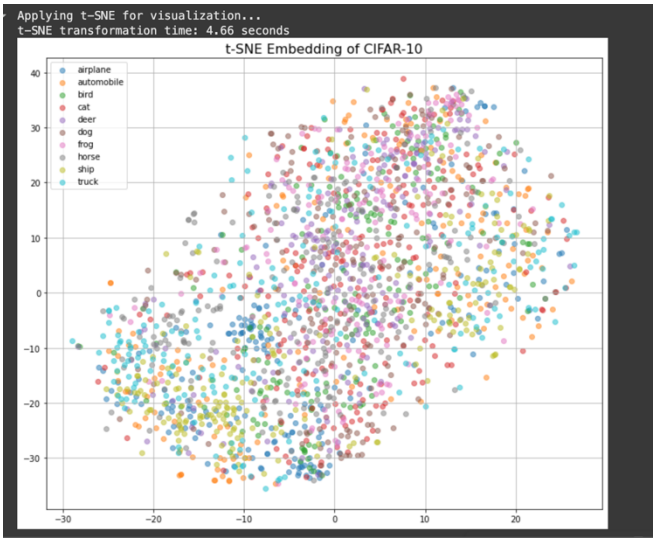
a) t-SNE Embedding

- Implementation:
  Applied StandardScaler followed by t-SNE (perplexity=30, random_state=42). Transformation completed in 4.66 seconds.

- Visual_Analysis:
  As shown , t-SNE produces tight clusters for "ship" (blue) and "truck" (orange), while "cat" (red) and "dog" (purple) overlap significantly. This explains the low F1-scores (~0.25) for these classes in prior classification tasks.

- Strengths: Effective at preserving local relationships (e.g., separating "airplane" from "bird").

- Limitations: Stochastic nature causes variability; runtime scales quadratically with sample size.



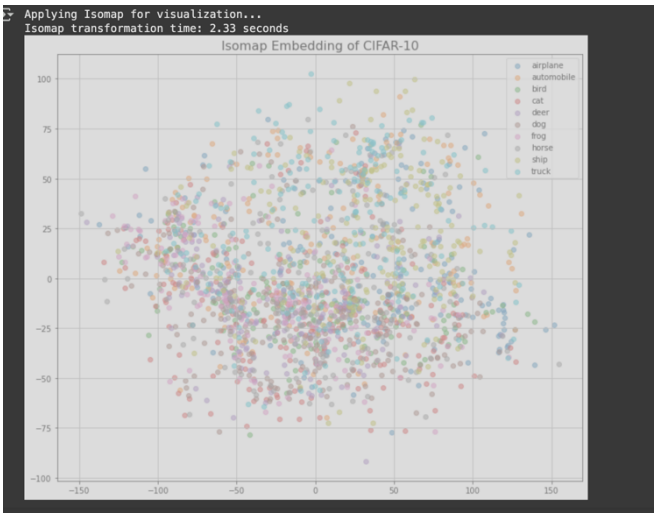
b) Isomap Embedding

- Implementation:
  Isomap with n_neighbors=10 achieved faster transformation (2.33 seconds).

- Visual_Analysis:
  It shows broader dispersion, with "automobile" (green) and "ship" (blue) forming distant clusters, while "deer" (cyan) and "horse" (gray) partially overlap. This suggests Isomap captures coarse semantic groupings (e.g., vehicles vs. animals).Strengths: Effective at preserving local relationships (e.g., separating "airplane" from "bird").

- Strengths: Efficient for global structure analysis; linear runtime with neighborhood size.

- Limitations: Sensitive to n_neighbors; loses fine-grained details (e.g., "bird" and "frog" intermixing).



IV. Conclusion & Obstacles

My investigations utilizing Random Forest (RF), Support Vector Classifier (SVC), and Logistic Regression (LR) on the CIFAR-10 dataset produced suboptimal outcomes, with accuracy rates ranging from 33.4% to 46.9%. Among the models assessed, the RF exhibited the highest performance, achieving an accuracy of 46.9% and an F1-score of 0.4649, attributable to its capacity to capture nonlinear relationships within the data. In contrast, the LR model demonstrated a lower accuracy of 33.4%, which can be attributed to the linear nature of pixel data. Furthermore, the application of manifold learning techniques, such as t-SNE and Isomap, indicated that the inadequate class separability in low-dimensional representations is directly associated with the misclassification of models, particularly for visually similar categories, such as "cat" and "dog."

The principal challenge identified in this research was the considerable computational expense associated with the training and evaluation of traditional machine learning models on the CIFAR-10 dataset. Although the Random Forest model attained the highest accuracy, it necessitated nearly three minutes for hyperparameter tuning, while the inference latency of the Support Vector Classifier (SVC) surpassed 80 seconds, presenting a significant limitation for practical application. These inefficiencies are attributed to the high dimensionality of image data and the intrinsic complexity of pixel-level feature extraction, which traditional methodologies struggle to process effectively. Moreover, manifold learning techniques such as t-SNE and Isomap, while providing valuable insights, were found to be computationally intensive even when applied to small data subsets, rendering them impractical for large-scale analyses without substantial optimization.

To mitigate these challenges, future research should emphasize contemporary deep learning methodologies that are more adept at handling image data. Convolutional Neural Networks (CNNs), including architectures such as ResNet or EfficientNet, have the potential to significantly enhance accuracy while utilizing GPU acceleration to decrease training duration. Furthermore, approximate manifold learning techniques like UMAP or hierarchical t-SNE may present more efficient alternatives for visualization, maintaining interpretability without incurring excessive computational costs. Another promising avenue is the implementation of transfer learning, wherein pretrained models fine-tuned on the CIFAR-10 dataset could eliminate the necessity for labor-intensive feature engineering while achieving state-of-the-art performance.

Ultimately, enhancing inference efficiency through strategies such as model pruning, quantization, and hardware-aware algorithm design could reconcile the disparity between accuracy and deployability. By transitioning towards deep learning and efficient algorithmic approaches, future investigations can address the computational limitations observed in this study while simultaneously improving generalization in complex image classification tasks.