ECEN 758 Project

Tharun Dubba
Computer Science Dept.
Texas A&M University
College Station, Texas
tharundtr@tamu.edu

Aakashdeep Sil
Computer Science Dept.
Texas A&M University
College Station, Texas
asil@tamu.edu

Sri Sai Kowshik Reddy Boyalla
Information and Operations Management Dept.
Texas A&M University
College Station, Texas
kowshik.boyalla@tamu.edu

Mithilesh Menakuru
Information and Operations Management Dept.
Texas A&M University
College Station, Texas
menakuru.mithilesh@tamu.edu

Kunpeng Wang
Electrical and Computer Engineering Dept.
Texas A&M University
College Station, Texas
kunpeng2023@tamu.edu

***Abstract*—This project focuses on classifying images from the CIFAR-10 dataset using various machine learning models, including logistic regression, decision trees, convolutional neural networks (CNNs), and transformers. The models were trained and evaluated to compare their performance using standard metrics like accuracy. Our results demonstrate the superior effectiveness of deep learning models, particularly CNNs and transformers, in handling complex image classification tasks.**

***Index Terms*—CIFAR-10, CNN, PCA, Transformers.**

I. Introduction

With the rapid development of artificial intelligence, visual data classification tasks play an important role in many fields, such as self-driving cars, industrial robots, and image searches in search engines. An efficient and accurate classification model can not only improve task efficiency but also greatly enhance the intelligence level of the system. In this project, we focus on the classic CIFAR-10 dataset. CIFAR-10 is a popular image dataset for image classification tasks. It contains 60,000 32x32 color images divided into 10 classes [1], offering a good balance between complexity and ease of experimentation, especially when computing resources are limited.

Logistic Regression and Decision Tree are both classic classification algorithms. Logistic Regression assumes a linear relationship between features and target values [2], while Decision Tree can capture nonlinear relationships in the data [3]. However, because the CIFAR-10 dataset has high-dimensional features and complex classes, simple models struggle to achieve satisfactory performance. Even when we tried training a Decision Tree with dimension-reduced data using PCA, the accuracy was still low. Our experiments show that these simple methods perform significantly worse than neural networks in image classification tasks.

To overcome these limitations, we implement the ResNet18 model [4], a deep residual network that uses skip connections to effectively address the gradient vanishing problem in deep networks. This allows it to capture more complex feature relationships. While ResNet18 achieved a significant improvement in performance on the CIFAR-10 classification task, our experiments revealed two problems: first, the model's accuracy was still not satisfactory; second, its generalization ability on zero-shot datasets was poor, indicating room for improvement in cross-domain adaptability.

Compared to ResNet18, Transformer models rely on the self-attention mechanism [5], which allows them to model global relationships between features more effectively. This makes them better suited for handling more complex learning tasks. Among Transformer-based models, the Compact Convolutional Transformer (CCT) combines the strengths of convolutional neural networks and Transformers [6] while significantly reducing computational costs through improvements to the Transformer structure [7]. We implement the CCT model, which further improves the performance of our classifier.

For the poor performance on zero-shot datasets, multimodal learning provided an effective solution [8]. By integrating information from images and text, the model's generalization ability can be improved. Therefore, we use ResNet18 as the image model and build a prompt dataset to create a text model. Through this multimodal architecture, we improve the model's performance on both the original test set and the zero-shot test set.

II. Method

A. Data Preparation

Our data preprocessing pipeline was critical in preparing the CIFAR-10 dataset for analysis and subsequent model training. Data cleansing, transformation, and splitting were implemented to ensure the dataset was optimized for machine learning workflows.

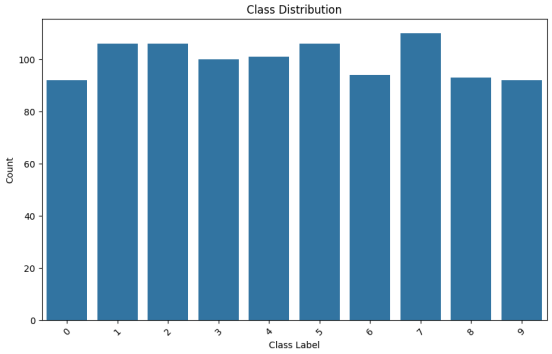
The CIFAR-10 dataset was imported using PyTorch's torchvision library. The dataset was organized into three subsets - Training Set (50,000 images), Validation Set (10,000 images) (split from the training set), and Test Set (10,000 images).

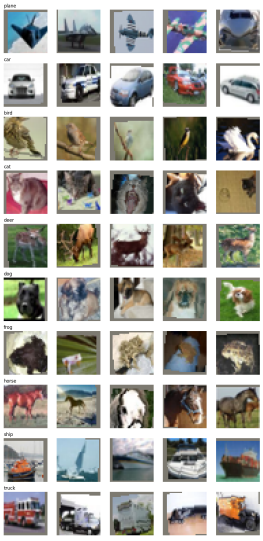
Each image in CIFAR-10 is a 32x32 pixel RGB image. To standardize the pixel values across all images, we calculated

the mean and standard deviation for each color channel. We got Mean (per channel): [0.4914, 0.4822, 0.4465], and Standard Deviation (per channel): [0.2470, 0.2435, 0.2616]. These values were used to normalize the images, ensuring that pixel intensities were scaled to have a mean of 0 and a variance of 1.

B. Exploratory Data Analysis

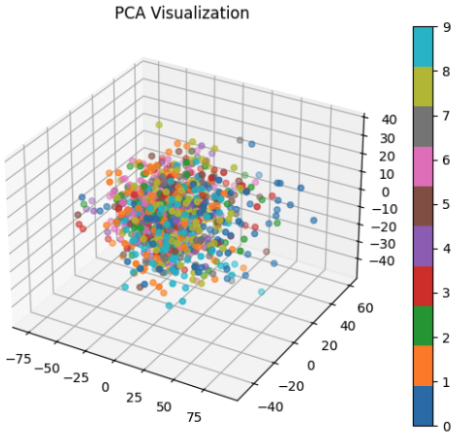
The analysis focuses on understanding the dataset's central tendencies, variability, distribution, and feature relationships to prepare it for machine learning. The basic statistical summary highlights key metrics, including count (to check for missing data), mean, median, standard deviation (for variability), and range (to identify outliers). The detailed summary adds metrics like skewness, kurtosis, IQR, and mode, providing deeper insights into distribution shape, normality, and outliers. The class distribution is for 1000 samples taken.



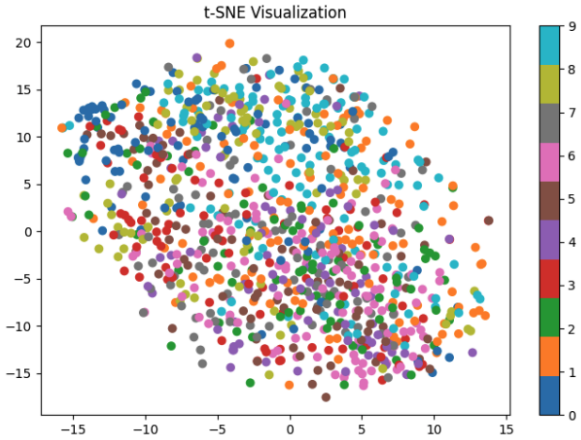


This visualization displays random examples from each class in the dataset. It helps in understanding the visual characteristics of each class, and verifying that the data is clean and correctly labeled.

Dimensionality reduction techniques are used to project high-dimensional data (e.g., pixel values) into a lower-dimensional space for visualization. These techniques help in understanding the structure of the data and identifying clusters or overlaps between classes.



PCA (Principal Component Analysis) reduces dimensions while retaining maximum variance. We visualized in 3D to show clusters of similar data points.



t-SNE (t-distributed Stochastic Neighbor Embedding) projects data into 2D or 3D space while preserving local relationships. It's effective for visualizing clusters but computationally expensive.

UMAP (Uniform Manifold Approximation and Projection) is similar to t-SNE but faster and better at preserving global structure. It produces clear separations between classes in 2D space.

Each point represents a sample, colored by its class label. It helps identify overlapping or distinct clusters of data points and is useful for datasets with high-dimensional features like images. These visualizations provide insights into the dataset's structure and separability, aiding in feature engineering and model selection.

C. Algorithm selection

Logistic Regression and Decision Trees are well-known foundational classification methods. Logistic regression is a simple and interpretable model that works well for linearly separable data. While it is not highly effective for complex datasets, it serves as a good baseline. Decision Trees offer more flexibility than logistic regression and can model non-linear relationships. They can handle mixed data and are resilient to outliers. Their performance on CIFAR-10 dataset can be limited to deep learning models.

To improve the performance further, we used neural network-based models. They can capture complex, nonlinear patterns in data, making them more suitable for tasks like image classification. We chose ResNet-18 because it offers a good balance between computational efficiency and classification accuracy among well-known model architectures.

To further improve the image recognition accuracy and the model's generalization ability, we explore two additional approaches.

The first approach is the Compact Convolutional Transformer (CCT). CCT effectively extracts both local and global features on the CIFAR-10 dataset by combining convolution operations with the self-attention mechanism. This design helps retain spatial features while capturing long-range dependencies, leading to better performance.

The second approach is multimodal learning. We combine the previously implemented ResNet-18 with a Bidirectional Long Short-Term Memory (BiLSTM) model to create a multimodal model that integrates visual and textual information. On the zero-shot test set, the multimodal learning method used text descriptions to supplement image features, helping the model make predictions even without training data for the target classes. This significantly improved the performance of zero-shot learning.

D. Model building

1) Machine Learning models:

Logistic regression - A linear model for binary and multiclass classification, logistic regression estimates the probabilities of class membership. This model is chosen for its interpretability, simplicity, and adaptability to multiclass classification, aligning with the diverse nature of our Cifar-10 categories. For Logistic Regression, we set the regularization parameter (C) to 1, employed automatic determination for the multi-class strategy, and used L2 penalty regularization.

Decision Trees - Decision Trees make decisions based on feature values to classify instances. This model is employed for its simplicity and interpretability, providing insights into the decision-making process for each classification. For Decision Trees, we set the criterion to "entropy," the maximum depth to 10, and used the "best" strategy for choosing splits.

2) ResNet-18:

ResNet-18 has 4 residual layers, each with 2 residual blocks. It uses skip connections to solve the gradient vanishing problem in deep networks. The structure can be shown with the following diagram [9]:

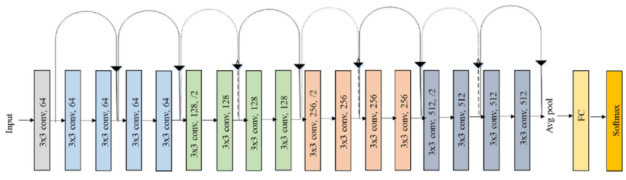


Fig. 1. ResNet-18 model

The ResNet-18 model can be loaded from PyTorch's built-in models. To fit the CIFAR-10 task with 10 classes, the last fully connected layer is modified. We first get the input feature size num_ftrs from the original layer and replace it with a new linear layer with an output size of 10. During training, we use the cross-entropy loss function $L = -\sum y_i \log(\hat{y}_i)$ as the objective and the Adam optimizer to adjust parameters. Each training step includes forward propagation to compute predictions and backward propagation to update the parameters.

3) CCT:

Compact Convolutional Transformers (CCT) build on the CVT architecture by introducing a convolutional tokenizer, which replaces the standard patch embedding approach with a convolutional block that uses ReLU activation and max pooling. This tokenizer captures local spatial information more effectively and flexibly, allowing for adjustable downsampling and token generation, potentially eliminating the need for positional embeddings to reduce computational costs. CCT also employs Sequence Pooling, a method that aggregates sequential embeddings from the transformer encoder, enabling the network to correlate data across input sequences while reducing computational complexity.

4) Multimodal model based on ResNet-18 and BiLSTM:

Based on ResNet-18, we built a multimodal model by adding a text model. The model takes both an image model and a text model as input, aligning their features through two linear layers. In our implementation, the image model loads pre-trained weights, and its parameters are frozen to speed up convergence. Specifically, image and text features are stacked into a 2D tensor, which is processed by the attention layer to capture relationships between the modalities. The fused features are passed through two fully connected layers: the first extracts high-level features and introduces non-linearity, while the second maps these features to the output space to generate predictions. The structure of this multimodal model is as follows:

Multimodal Model

Input: Image (batch_size, channels, height, width), Text (batch_size, sequence_length)
Layer 1: Process image through pretrained model: (batch_size, image_feature_dim)
Layer 2: Process text through pretrained model: (batch_size, text_feature_dim)
Layer 3: Apply linear transformation to image features: (batch_size, 128)
Layer 4: Apply linear transformation to text features: (batch_size, 128)
Layer 5: Apply multihead attention:(batch_size, 2, 128)
Layer 6: Reshape attention output: (batch_size, 256)
Layer 7: Apply linear transformation: (batch_size, 64), ReLU activation
Layer 8: Apply linear transformation: (batch_size, 10)
Output: Class probabilities

Specifically, in the text model, we use a bidirectional LSTM to capture both forward and backward context in the text, which is widely used in natural language processing. After

averaging the LSTM output, it is classified through a fully connected layer, with dropout applied to prevent overfitting.

During training, the loss function and optimizer are the same as in ResNet-18. The difference is that both image and text data are fed into the model for forward and backward propagation to update the parameters.

E. Hyperparameter tuning

The hyperparameter tuning process for Decision trees involved a exploration of maximum depth to optimize its performance. We tested it on different depths, ranging from 1-10 and without any depth. We found out that a depth of 10 gave the best accuracy. For ResNet18, we explored the choice of optimizer, experimenting with both Adam (Adaptive Moment Estimation) and SGD (Stochastic Gradient Descent). The learning rate plays a crucial role in determining the step size during optimization. We identified the combination of components to identify the best accuracy and convergence time. Fine-tuning CCT model involved multiple steps, we tested with different lengths of residual connections and learning rates, for which we used a scheduler.

F. Model evaluation

In the exploration of the CIFAR-10 dataset, the selection of best models involved a meticulous process where we considered ML models, CNN, and CCT as candidate models. The CNNs excelled in feature extraction and hierarchical pattern recognition, showcasing superior performance compared to the other models. This efficiency arises from the specific architecture and operations performed by CNNs, which leverage shared weights and hierarchical feature extraction. The selection of the best model was ultimately driven by a comprehensive evaluation of accuracy, computational efficiency, and the model's ability to generalize well to unseen data, with the CCT model emerging as the optimal choice.

III. Experimental Results and Discussion

A. Experimental Results

TABLE I
Model and Corresponding Accuracy on Test Data

Test Type	Accuracy (%)	Precision	Recall	F1-score
Logistic Regression	40	0.39	0.40	0.39
Decision Tree	31	0.31	0.31	0.31
Decision Tree with PCA	31	0.31	0.31	0.31
ResNet18	80.64	0.80	0.80	0.80
CCT	94.4	0.94	0.94	0.94

The training and experiments of the multimodal model are unique. We selected seven classes from CIFAR-10 as the training set, ensuring the model has not seen the other three classes for zero-shot testing. We tested both ResNet-18 and the multimodal model using the in-domain and zero-shot test sets, and the results are as follows: By comparing the performance of both models, we can clearly see how the multimodal improvement boosts ResNet-18's performance. The performance improvement on the regular test set shows that

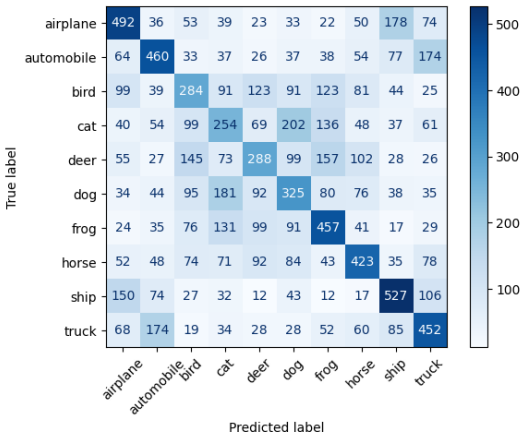


Fig. 2. Logistic Regression confusion matrix

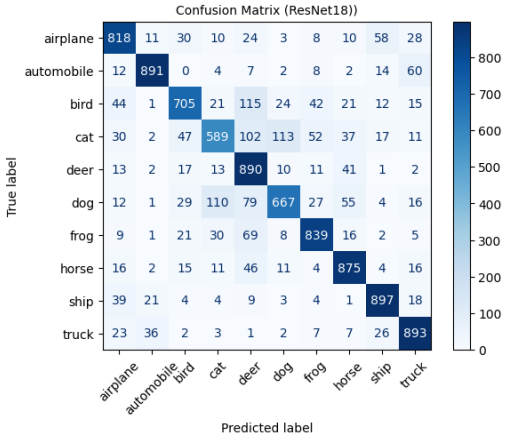


Fig. 3. ResNet18 confusion matrix

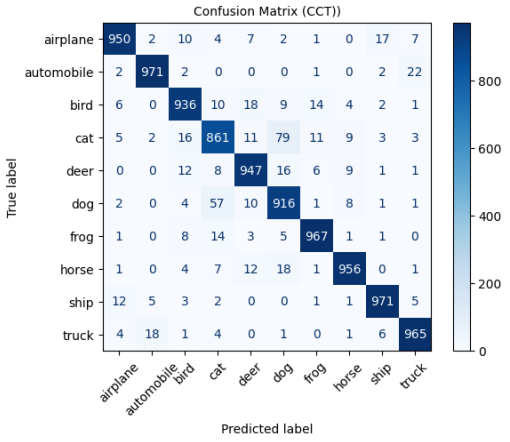


Fig. 4. CCT confusion matrix

TABLE II
MULTIMODAL ZEROSHOT EXPERIMENT

Test Type	Accuracy (%)	Precision	Recall	F1-score
ResNet18 In-domain	80.11	0.85	0.80	0.82
ResNet18 Zero-shot	0.00	0.00	0.00	0.00
Multimodal In-domain	85.40	0.85	0.85	0.85
Multimodal Zero-shot	33.02	0.33	0.33	0.33

combining images and text helps enhance recognition, which matches our expectations. ResNet-18 was unable to recognize classes not seen during training, but with simple multimodal improvements, it gained this ability. With further improvement in text dataset richness and computational resources, this multimodal model can show even stronger performance.

B. Model interpretablity

We can experiment with model interpretability by calculating the input image gradients and visualizing the contribution of input features using a heatmap. The image below shows how ResNet-18 is influenced by the input features. This visualization aids in understanding which parts of the input image are crucial for the network's classification decisions. Although interpreting Transformer models is still very difficult, there is an ongoing research on finding methods to better understand the reasoning. The visualization shows that the model focuses

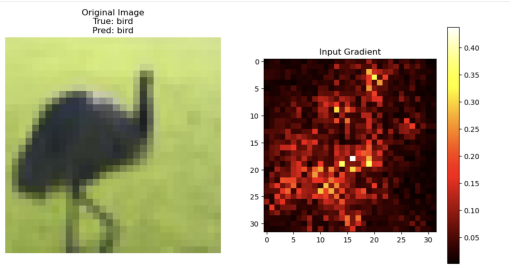


Fig. 5. Model interpretability visualization

on key areas (light regions) such as the bird's head, body, and legs, indicating that the model successfully captures these important features for classification.

C. Business insights

This project achieved over 90% accuracy and improved traditional models with multimodal learning, both of which have significant commercial value.

CCT's high accuracy in image recognition shows great potential. Its global feature modeling capability allows it to perform well not only on simple datasets like CIFAR-10 but is expected to also on more complex datasets. Additionally, compared to other Transformer models, CCT's low computational cost makes it suitable for deployment on edge devices or embedded systems with limited resources. In industrial production, CCT can quickly classify complex image data with low computational requirements, enabling tasks such as quality inspection and product sorting.

The multimodal model can be applied to scenarios involving both text and image data, especially where zero-shot classification is common. For example, in autonomous driving, it can recognize new traffic signs based on traffic rules. For e-commerce platforms, it can quickly identify and recommend new products. In healthcare, it can combine medical images with electronic health records to improve the diagnosis of rare conditions.

IV. CONCLUSION

In conclusion, our experiment with the CIFAR-10 dataset employing a different models spanning traditional machine learning techniques, Convolutional Neural Networks (CNN) and Transformers has provided valuable insights into their respective capabilities for image classification. Traditional models, including Logistic regression, demonstrated competitive performance, although it is constrained with only linearly separable data. Decision tree classifier didn't do well, while the CNN, specifically tailored for image-based tasks, outperformed ML models, showcasing its effectiveness in discerning nuanced features inherent in the images. With the recent rise in Transformers applications, accuracy has risen significantly. CCT model provided a great trade off between the size of the model and accuracy. This comprehensive evaluation underscores the strengths of each model class, providing a foundation for informed model selection in the context of image classification tasks.

REFERENCES

[1] Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. (2009).

[2] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–242.

[3] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.

[4] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan & R. Garnett (eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA (p./pp. 5998–6008), .

[6] Hassani, A., & Walton, S. (2021). Escaping the Big Data Paradigm with Compact Transformers. arXiv preprint arXiv:2104.05704.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations, .

[8] Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.

[9] Ramzan F, Khan MUG, Rehmat A, Iqbal S, Saba T, Rehman A, Mehmood Z. A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks. J Med Syst. 2019 Dec 18;44(2):37.