COMP 4447 Final Project

1. Introduction

Object recognition is a very popular topic in the computer vision. The mainstream approach is to implement deep learning-based models, such as Convolutional Neural Networks, Transformers, and so on, which provides very accurate prediction. Very few work still insist on using traditional machine learning models, like Support Vector Machine and Random Forest. However, traditional machine learning models are more interpretable than deep learning ones.

1.1 Problem Statement

The traditional ones learn the data representative or patterns from the sparse data distribution, while the deep learning ones require a large number of samples to acquire the data feature well. To some degree, learning the dense distribution equals to study the data's manifold. Therefore, the traditional models perform stably by costing much less data samples and computational source. In the other side, in realistic, the scale of dataset is usually small like hundred- or thousand-level instead of being sufficient like million-level, which is more suitable for utilizing traditional model. These two reasons supports that the traditional ones are still valuable to research.

1.2 Research Direction

This project is to study if traditional models can learn image feature and recognize object well, and to research the drawbacks that prevent the traditional models behave as good as the deep learning ones. The common studies focus more on the advantage of deep learning models over the traditional one. Our study pays attention to the inversed version and providing new perspective on potentially improving traditional models.

2.Model

We implemented Support Vector Classifier (SVC) and Random Forest Classifier (RFC) to classify images.

2.1 SVC

The Support Vector Classifier (SVC) identifies an optimal hyperplane in a high-dimensional feature space to maximize the margin between separable classes. For non-linearly separable data, SVC employs kernel functions (e.g., Radial Basis Function, polynomial) to implicitly map inputs into higher-dimensional spaces where linear separation becomes feasible. The optimization objective minimizes:

$$\frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{n} \xi_i$$

where **w** defines the hyperplane, $C>0$ is a regularization parameter penalizing misclassifications, and ξi are slack variables accommodating margin violations. SVC exhibits strong generalization performance, robustness to high-dimensional data, and adaptability to complex decision boundaries via kernel methods. Its computational complexity scales with the number of support vectors, making it efficient for mid-sized datasets.

2.2 RFC

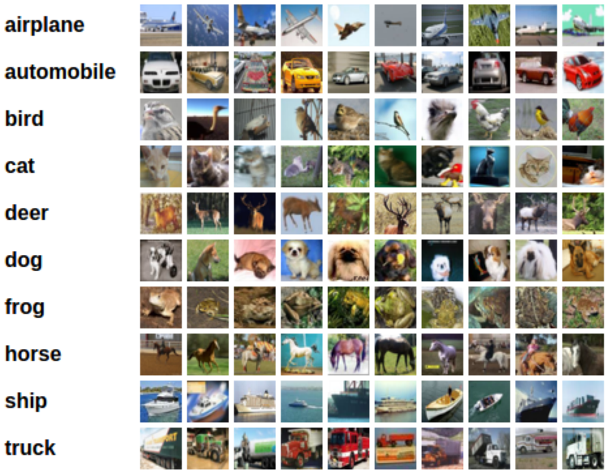
Random Forest Classifier (RFC) is an ensemble learning method that constructs multiple decision trees during training and aggregates predictions via majority voting (classification). Each tree is trained on a bootstrap sample of the dataset (bagging) with random feature subsets at each split. This dual randomness—data-level (bagging) and feature-level (random subspaces)—decorrelates individual trees, enhancing robustness against overfitting and noise. In details, RFC follows three steps, bootstrap sampling for each tree, randomized construct each tree (maximizes information gain or entropy to split each node), majority voting for prediction.

3.Experiment

3.1 Dataset

In this work, the selected dataset is CIFAR-10 dataset (Canadian Institute For Advanced Research, 10 classes), a canonical, extensively curated benchmark dataset within the computer vision and machine learning research communities. CIFAR-10 comprises 60,000 color images, 50,000 training images, 10,000 testing ones, distributed across 10 mutually exclusive object classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck). Each class contains precisely 6,000 images. The size of each image is [32,32,3].

The right figure show image samples of 10 classes in the CIFAR-10.



3.2 Data Preprocessing

To improve the robustness of the model, we augmented and normalized the images. The image augmentation methods include adding random noise, randomly rotating the image between -20 and 20 degrees, horizontal flipping, randomly adding gaussian noise.

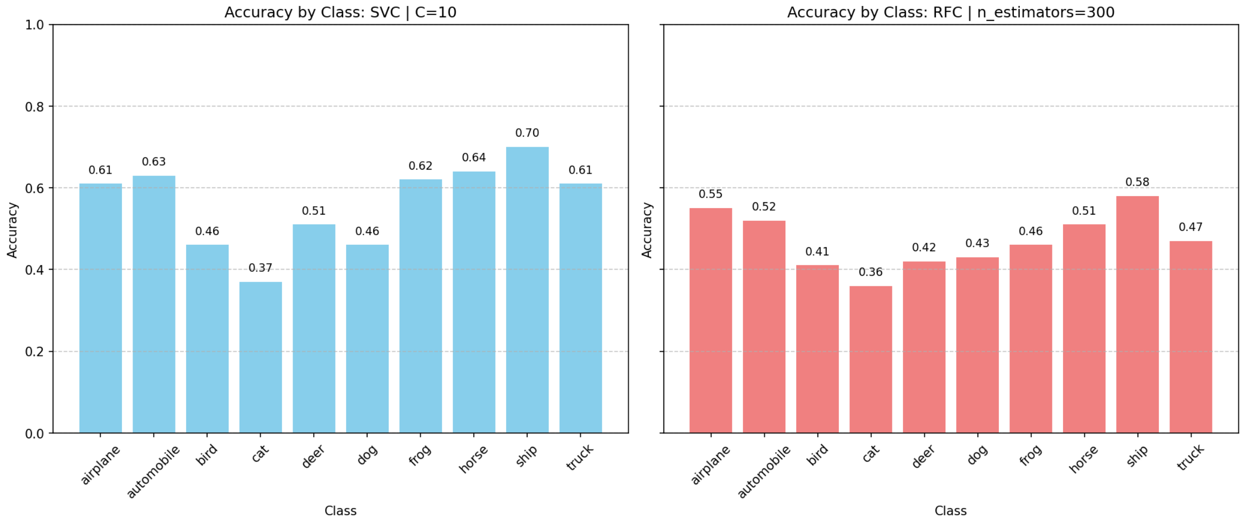
Then, we flattened the images from [32,32,3] to [1, 32*32*3], and normalized them by z-score normalization.

3.3 Experimental Design

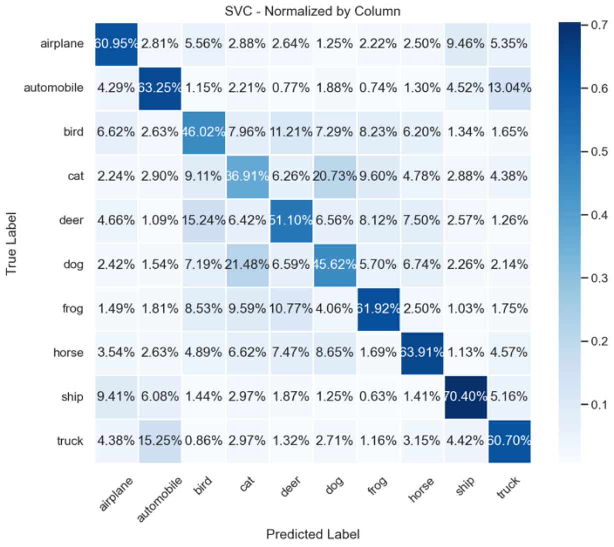
We separately train SVC and RFC on the CIFAR-10 dataset by using the CPU embedded in the Apple Macbook Air. The hyperparameter C of SVC is 10, the number of estimator of RFC is 300.

3.4 Experimental Results

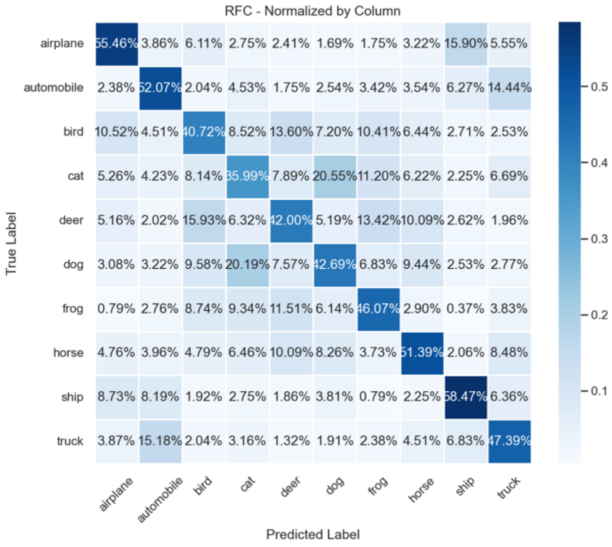
For overall accuracy across all 10 categories, SVC achieved 56.03% accuracy, around 10% higher than RFC's result, 47.12%. Hence, SVC performs better than RFC on the CIFAR-10 dataset. The bellowed figures support this view as well.



In addition, both SVC and RFC predict ship better than the other categories. Furthermore, vehicle is easier to predict than animals. For instance, in SVC, the accuracy of ship, automobile, airplane, and truck are generally higher than that of six animal categories. There is similar pattern in the result of RFC as well. This heatmap upholds this view as well.



	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	60.95%	2.81%	5.56%	2.88%	2.64%	1.25%	2.22%	2.50%	9.46%	5.35%
automobile	4.29%	63.25%	1.15%	2.21%	0.77%	1.88%	0.74%	1.30%	4.52%	13.04%
bird	6.62%	2.63%	46.02%	7.96%	11.21%	7.29%	8.23%	6.20%	1.34%	1.65%
cat	2.24%	2.90%	9.11%	36.91%	6.26%	20.73%	9.60%	4.78%	2.88%	4.38%
deer	4.66%	1.09%	15.24%	6.42%	51.10%	6.56%	8.12%	7.50%	2.57%	1.26%
dog	2.42%	1.54%	7.19%	21.48%	6.59%	45.62%	5.70%	6.74%	2.26%	2.14%
frog	1.49%	1.81%	8.53%	9.59%	10.77%	4.06%	61.92%	2.50%	1.03%	1.75%
horse	3.54%	2.63%	4.89%	6.62%	7.47%	8.65%	1.69%	63.91%	1.13%	4.57%
ship	9.41%	6.08%	1.44%	2.97%	1.87%	1.25%	0.63%	1.41%	70.40%	5.16%
truck	4.38%	15.25%	0.86%	2.97%	1.32%	2.71%	1.16%	3.15%	4.42%	60.70%



	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	55.46%	3.86%	6.11%	2.75%	2.41%	1.69%	1.75%	3.22%	15.90%	5.55%
automobile	2.38%	52.07%	2.04%	4.53%	1.75%	2.54%	3.42%	3.54%	6.27%	14.44%
bird	10.52%	4.51%	40.72%	8.52%	13.60%	7.20%	10.41%	6.44%	2.71%	2.53%
cat	5.26%	4.23%	8.14%	35.99%	7.89%	20.55%	11.20%	6.22%	2.25%	6.69%
deer	5.16%	2.02%	15.93%	6.32%	42.00%	5.19%	13.42%	10.09%	2.62%	1.96%
dog	3.08%	3.22%	9.58%	20.19%	7.57%	42.69%	6.83%	9.44%	2.53%	2.77%
frog	0.79%	2.76%	8.74%	9.34%	11.51%	6.14%	46.07%	2.90%	0.37%	3.83%
horse	4.76%	3.96%	4.79%	6.46%	10.09%	8.26%	3.73%	51.39%	2.06%	8.48%
ship	8.73%	8.19%	1.92%	2.75%	1.86%	3.81%	0.79%	2.25%	58.47%	6.36%
truck	3.87%	15.18%	2.04%	3.16%	1.32%	1.91%	2.38%	4.51%	6.83%	47.39%

However, SVC's performance is far from competitive than deep learning models'. The following table shows that ViT-H/14 (Dosovitskiy, 2020) and DINOv2 (Oquab, 2023) are very powerful and close to perfect on Cifar-10.

Model	**Accuracy**
ViT-H/14 (Dosovitskiy, 2020)	99.50%,
DINOv2 (Oquab, 2023)	99.50%,
SVC (Ours)	56.03%
RFC (Ours)	47.12%

3.5 Ablation Study

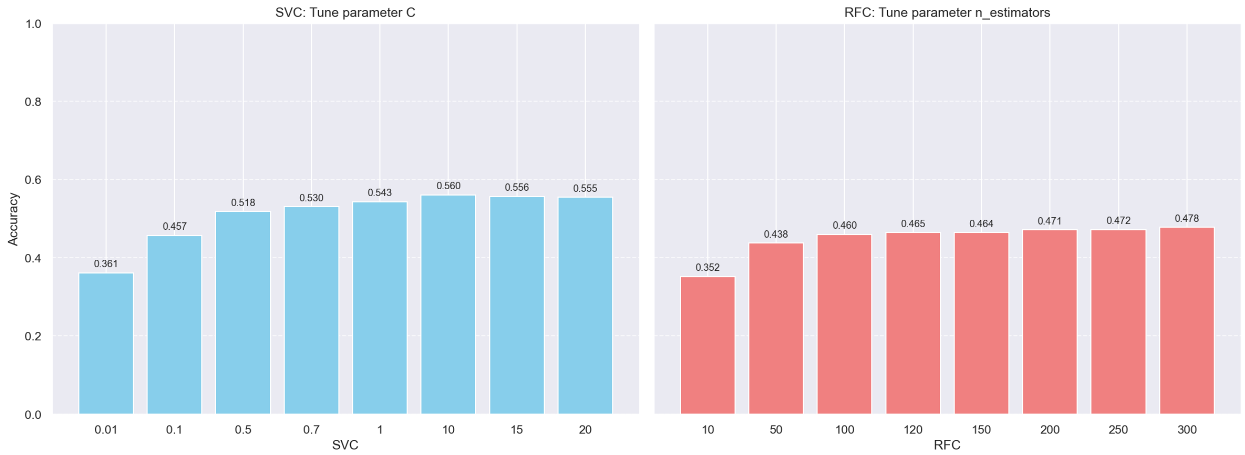
To push the limitation of SVC and RFC, we tuned their hyperparameters to check the differences.

3.5.1 Study the effectiveness of C on SVC performance

We set C = [0.01,0.1,0.5,0.7,1,10, 15] and trained SVC one by one. The following figure shows the corresponding accuracy is [0.3612, 0.4569, 0.5179, 0.5305, 0.5434, 0.5603, 0.5564], indicating that larger C benefits the better prediction results. When C=10, SVC reaches the best performance.

3.5.2 Study the effectiveness of number of estimators on RFC performance

We set n_estimators = [10,50,100,120,150,200,250,300] and trained RFC one by one. The following figure shows the corresponding accuracy is [0.3517, 0.4375, 0.4595, 0.4645, 0.4643, 0.4712, 0.472, 0.478] indicating that larger n_estimators benefits the better prediction results. When n_estimators=300, RFC reaches the best performance.



4.Discussion

This project aims to learn if traditional models like SVC and RFC can finish image classification as good as deep learning ones. This experimental result indicates that SVC performs better than RFC. When C equals to 10, SVC achieved its best accuracy 56.03%, around 10% higher than RFC's best accuracy, 47.2%. However, the state of the art (SOTA) result of CIFAR-10 is 99.5%, coming from ViT-H/14 and DINOv2. Morevoer, ViT-H/14 and DINOv2 have 632M and 1.1B tunable parameters, while SVC and RFC possess 5 (C, kernel, gamma, etc) and 10+ (n_estimators, max_depth, max_features, min_samples_split, min_samples_leaf, etc.) tunable parameters. This reveals that ViT-H/14 and DINOv2 have much more complicate structure and can extract much richer features than SVC and RFC. With simple structure, SVC and RFC can only extract rare features, impossible for reaching extremely good performance.

It is the richness of feature that deciding the difference of traditional models and deep learning ones. From another perspectives, traditional models such as SVC and RFC are very efficient in detection with merely few tunable parameters. In the meanwhile, many parameters in the deep learning models may be redundant. Thus, traditional models is weak to recognize object well out of image data due to the incapability of capturing sufficient and various features, but they learn image features efficiently so that they can still output stable performance while facing with complex task like object recognition.

5.Conclusion

Traditional models like SVC and RFC extract insufficient features from image dataset, causing the mediocre performance in the object recognition. The experiment on the CIFAR-10 dataset verifies this statement. Simultaneously, their stable performance also indicates the efficiency of the features they extracted. The efficiency is what deep learning models urgently need. In the context of artificial intelligence, efficiency means to fuse physics information, mathematical equations, or chemistry equations into the neural network. We can take them as supplementary features and design related loss function to control the convergence. What's more, we can even implement them while computing gradients and update the gradient, which is much helpful. Hence, in the future, improving the efficiency of the learned representative is a good research direction for deep learning models. Conversely, elevating the variant of feature is the primary task of traditional models, which pushes them to be close to the format of deep learning models, and detaches them from their name, traditional models. Perhaps, concentrating on the structured data instead of complex data modality like video, acoustic, and images is the direction of traditional models. Maintaining stable performance on the few-shot dataset is also a good option.

Reference

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.