TERAFAC SUBMISSION REPORT

Dataset used : Stanford Cars Dataset

LAKSHAY VERMA
AIML B2
11619051622
USAR

Abstract

Fine-grained image classification is a challenging computer vision problem because different classes often share very similar visual characteristics. In this project, we address a fine-grained vehicle classification task using the Stanford Cars196 dataset as part of the Terafac Image Classification Challenge.

Instead of focusing only on achieving high accuracy, this work emphasizes building a clean and well-reasoned training pipeline, understanding model behavior, and analyzing results at each stage. The approach progresses systematically from a baseline transfer learning model to more advanced techniques, including data augmentation, architectural fine-tuning, interpretability analysis, and ensemble learning.

EfficientNet-B4 was first used as a strong baseline model, followed by training improvements and detailed analysis of model attention using Grad-CAM visualizations. To further improve robustness, an ensemble of EfficientNet-B4 and ConvNeXt-Tiny was constructed using soft-voting of prediction probabilities. This ensemble achieved a validation accuracy of approximately 92.33%, demonstrating improved generalization compared to individual models.

Throughout the project, careful attention was given to dataset splitting, reproducibility, and honest evaluation of failures and limitations. The goal of this work is not only to present a working solution, but also to demonstrate a thoughtful and structured approach to solving real-world computer vision problems.

Introduction

Image classification is a fundamental problem in computer vision, but its difficulty increases significantly in fine-grained settings. In fine-grained classification tasks, different classes often look very similar to each other, while variations within the same class can still exist due to changes in viewpoint, lighting, and background. The Stanford Cars196 dataset is a clear example of this challenge, where different car models may differ only in small visual details such as headlight shape, grill design, or logo placement.

Unlike generic image classification problems, fine-grained vehicle classification requires the model to focus on subtle and localized features rather than global object appearance. This makes the task more sensitive to background noise, dataset bias, and overfitting. Simply increasing model size or training for more epochs does not always lead to meaningful improvements and can often result in poor generalization.

The Terafac Image Classification Challenge emphasizes not only model performance, but also the reasoning behind architectural choices, training strategies, and result analysis. With this perspective, the goal of this project was approached in a structured and incremental manner. Instead of directly applying complex methods, the work was divided into multiple levels, where each level builds on the understanding gained from the previous one.

The project begins with a strong but simple baseline using transfer learning, ensuring that the data pipeline, dataset splits, and evaluation setup are correct. Subsequent levels introduce controlled improvements such as data augmentation, regularization, and optimization strategies. More advanced stages focus on understanding model behavior through interpretability techniques and improving robustness using ensemble learning.

Throughout this work, emphasis was placed on asking *why* a particular method should work before applying it, and *what* the model is actually learning after training. By following this approach, the project aims to demonstrate not just a working solution, but a thoughtful problem-solving process that reflects real-world machine learning development.

Dataset & Challenges

This project uses the **Stanford Cars196** dataset, which is a widely used benchmark for fine-grained vehicle classification. The dataset contains images of cars belonging to **196 different classes**, where each class represents a specific car make, model, and year. The dataset provides both class labels and bounding box annotations for each image.

Dataset Characteristics

The dataset is divided into official training and test splits, and includes:

- High-resolution images captured in real-world conditions
- Significant visual similarity between many classes
- Variations in viewpoint, lighting, background, and scale
- Bounding box annotations that localize the car within each image

Although the dataset is relatively clean compared to web-scraped datasets, it still presents challenges such as subtle label differences and limited visual cues for certain classes.

Fine-Grained Nature of the Problem

The primary difficulty of the Cars196 dataset lies in its **fine-grained structure**. Many classes differ only in very small visual details, such as:

- Headlight or tail-light shape
- Grill patterns
- Logo placement
- Minor changes in body design across model years

At the same time, cars from different classes can appear almost identical at a global level. This makes it difficult for models that rely only on coarse features or background context to perform well.

Background Noise and Localization

Another challenge is the presence of complex and varied backgrounds. Images are often captured in outdoor environments with roads, buildings, trees, or other vehicles. If the model learns to associate background patterns with specific classes, it can lead to poor generalization.

To address this, the provided **bounding box annotations** were used to crop the vehicle region before training. This helps the model focus on the car itself rather than irrelevant background information, which is especially important in fine-grained classification tasks.

Dataset Split and Class Balance

Following the Terafac challenge guidelines, a strict **80-10-10 split** was maintained for training, validation, and testing. The validation set was derived from the training data using **stratified sampling**, ensuring that all classes are represented proportionally.

Maintaining class balance during splitting is important in this dataset because some classes have fewer examples than others. Without stratification, evaluation metrics could become misleading.

Summary of Challenges

In summary, the key challenges posed by the Cars196 dataset are:

- Extremely high inter-class similarity
- Subtle discriminative features
- Sensitivity to background noise
- Risk of overfitting when using large models
- Need for careful dataset splitting and evaluation

These challenges motivated the use of transfer learning, careful preprocessing, interpretability analysis, and ensemble methods in later stages of the project.

Methodology

The methodology for this project was designed in a progressive manner, where each stage builds upon the understanding gained from the previous one. Instead of directly applying complex techniques, the approach was to start simple, verify correctness, and then introduce additional complexity only when it was justified by observed behavior.

The implementation was organized according to the multi-level structure defined in the Terafac challenge, with each level serving a specific purpose.

3.1 Level 1 – Baseline Model

The first step was to establish a reliable baseline using transfer learning. For this purpose, **EfficientNet-B4**, pretrained on ImageNet, was selected due to its strong balance between accuracy and computational efficiency.

The baseline model was trained on cropped vehicle images using the provided bounding box annotations. This ensured that the model focused primarily on the car rather than the surrounding background. At this stage, the goal was not aggressive optimization, but rather to verify that:

- The dataset was loaded correctly
- The train, validation, and test splits were correct
- The training loop and evaluation metrics were functioning as expected

This baseline provided a strong reference point for evaluating later improvements.

3.2 Level 2 – Training Improvements

After establishing a stable baseline, the next step was to improve generalization through controlled training enhancements. Several commonly used techniques were applied, including data augmentation and regularization.

Data augmentation such as horizontal flipping and color jittering was introduced to help the model become invariant to minor appearance changes. Label smoothing was used in the loss function to reduce overconfidence in predictions, which is particularly useful in fine-grained classification tasks where class boundaries are subtle.

The optimizer was changed to AdamW, and a cosine annealing learning rate scheduler was used to provide smoother convergence. These changes were applied incrementally, with validation performance monitored closely to ensure that improvements were genuine and not caused by overfitting.

3.3 Level 3 – Advanced Architecture and Interpretability

At this stage, the focus shifted from purely improving accuracy to understanding **how** the model makes decisions. The EfficientNet-based model was fine-tuned more deeply, allowing the pretrained backbone to adapt to the specific visual characteristics of car models.

To interpret the model's behavior, **Grad-CAM** was used to visualize the regions of the image that contributed most to the model's predictions. These visualizations provided valuable insight into whether the model was attending to meaningful vehicle parts such as headlights, grills, and wheels, or whether it was relying on irrelevant background cues.

The Grad-CAM analysis helped validate that the model was learning reasonable and explainable features. It also revealed cases where the model struggled, particularly when two classes were visually almost indistinguishable even within the cropped region.

3.4 Level 4 – Ensemble Learning

The final stage focused on improving robustness rather than adding architectural complexity. Instead of using meta-learning or reinforcement learning, **ensemble learning** was chosen because it is simple, effective, and widely used in practical systems.

Two independently trained models were used:

- EfficientNet-B4
- ConvNeXt-Tiny

These models differ significantly in architecture and feature representation, which makes them suitable candidates for an ensemble. During inference, both models produced class probability distributions for each input image. These probabilities were then averaged using a **soft-voting strategy**, and the final prediction was obtained by selecting the class with the highest averaged probability.

The motivation behind this approach was to reduce model-specific biases and errors. Even when individual models make different mistakes, combining their predictions can lead to more stable and reliable performance.

3.5 Summary of Methodological Choices

Overall, the methodology emphasizes:

- Incremental development rather than abrupt complexity
- Validation-driven decision making
- Preference for simple, well-understood techniques
- Focus on interpretability and robustness

This approach reflects a real-world machine learning workflow, where understanding and reliability are often more important than marginal gains in accuracy.

Experiments & Results

The experimental evaluation in this project was conducted with the primary goal of understanding model behavior rather than aggressively maximizing numerical performance. All experiments followed a consistent training and evaluation protocol to ensure that observed improvements were meaningful and reproducible.

Training Setup

All models were trained using transfer learning with ImageNet-pretrained weights. Input images were cropped using the provided bounding box annotations to reduce background influence. Training was performed using the AdamW optimizer with a cosine annealing learning rate schedule. Label smoothing was applied to reduce overconfidence, especially in cases where class boundaries were visually ambiguous.

Validation performance was monitored closely during training, and model checkpoints were saved based on validation accuracy to avoid overfitting.

Baseline Performance

The baseline EfficientNet-B4 model provided a strong starting point, achieving high validation accuracy without extensive tuning. This confirmed that the data pipeline, dataset splits, and preprocessing steps were correctly implemented. The baseline also highlighted the inherent difficulty of the dataset, as further gains required more careful reasoning rather than simple architectural scaling.

Effect of Training Improvements

Introducing data augmentation and regularization techniques led to more stable training and improved validation performance. Although these changes did not drastically increase peak accuracy, they reduced fluctuations across epochs and improved generalization. This observation reinforced the importance of training discipline over brute-force optimization.

Advanced Fine-Tuning and Interpretability

Deeper fine-tuning of the EfficientNet model improved its ability to capture fine-grained visual details. Grad-CAM visualizations showed that the model focused on meaningful regions of the vehicle, such as headlights, grills, and body contours, rather than irrelevant background areas. This provided confidence that the model’s predictions were based on valid visual cues.

However, some misclassifications persisted, particularly between car models that differed only in subtle design elements. These cases demonstrated the practical limits of single-model performance on fine-grained datasets.

Ensemble Results

To improve robustness, an ensemble of EfficientNet-B4 and ConvNeXt-Tiny was constructed using soft voting. The ensemble achieved a validation accuracy of approximately **92.33%**, which was comparable to or slightly better than the individual models. While the numerical improvement was modest, the ensemble showed more consistent behavior across different classes.

The ensemble results suggest that combining models with different architectural biases can reduce model-specific errors, even when overall accuracy gains are small. This trade-off between complexity and robustness was considered acceptable for the goals of this project.

Summary of Experimental Findings

Overall, the experiments indicate that:

- Strong baselines are crucial in fine-grained classification tasks
- Careful training strategies can improve stability more than raw accuracy
- Interpretability tools are valuable for validating learned representations
- Ensemble methods improve robustness even with limited gains in accuracy

Error Analysis

Despite achieving strong overall performance, the models still exhibit errors that highlight the inherent difficulty of fine-grained vehicle classification. Analyzing these errors was an important step in understanding the limitations of the approach and avoiding misleading conclusions based solely on accuracy values.

A common source of misclassification occurs between car models that are visually almost identical. In many cases, different classes differ only in very small design changes such as minor variations in headlight shape, grill pattern, or logo placement. When such features are subtle or partially occluded, even a well-trained model struggles to distinguish between classes reliably.

Another observed issue arises from viewpoint and lighting variations. Some images are captured from unusual angles or under poor lighting conditions, which can hide discriminative features. Although bounding box cropping helps reduce background noise, it cannot fully compensate for missing or unclear visual cues within the vehicle region itself.

In a few cases, misclassifications appear to be influenced by intra-class variability. Cars from the same class may look different due to color, accessories, or modifications, which can confuse the model when these variations resemble features from other classes. This suggests that visual diversity within classes remains a challenge, even when using pretrained models.

Importantly, the error analysis indicates that most failures are not random but occur in predictable and explainable scenarios. This supports the conclusion that the models are learning meaningful representations, even though certain fine-grained distinctions remain difficult. These observations also justify the use of ensemble methods, which help reduce model-specific errors but cannot completely eliminate ambiguity inherent in the dataset.

Insights & Learnings

This project provided several practical insights into fine-grained image classification and real-world model development. One key learning was that having a clean and correct data pipeline is more important than applying complex techniques early. Small mistakes in dataset splitting or preprocessing can significantly affect results.

Another important insight was that increasing model complexity does not always lead to meaningful improvements. Strong pretrained models already capture powerful representations, and further gains often require careful reasoning rather than aggressive tuning. Bounding box cropping proved to be a simple but effective step in helping the model focus on relevant visual features.

Using interpretability tools such as Grad-CAM helped build confidence in the model's behavior. Visualizing attention regions made it easier to understand both correct predictions and failure cases, and confirmed that the model was not relying on irrelevant background cues.

Finally, ensemble learning demonstrated that combining diverse models can improve robustness even when accuracy gains are small. Overall, this project reinforced the importance of thinking carefully before applying techniques and validating each step through observation and analysis.

Conclusion

In this work, a structured approach was followed to solve a fine-grained image classification problem using the Stanford Cars196 dataset. Instead of directly applying complex methods, the solution was built incrementally, starting from a strong baseline and gradually incorporating improvements based on observed results and analysis.

Through careful preprocessing, transfer learning, interpretability analysis, and ensemble learning, a robust and reproducible system was developed. While the final ensemble achieved competitive performance, equal importance was given to understanding model behavior, analyzing errors, and acknowledging limitations.

Overall, this project demonstrates a thoughtful and disciplined approach to machine learning problem-solving, where reasoning, experimentation, and reflection guide each decision. The insights gained from this process are valuable not only for this specific task, but also for tackling similar real-world computer vision problems in the future.