Advancing Fine Grained Recognition Through Weakly Supervised Learning

Ranadeep Mahendra Prathapagiri
ranadeep@buffalo.edu
50538628

Sri Charan Byreddy
sbyreddy@buffalo.edu
50534329

***Abstract*—This project aims to enhance fine-grain recognition by iteratively refining models and comparing their performance. Starting with a basic model using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM), we explore advanced techniques, including replacing SVM with a Multi-Layer Perceptron (MLP). Our goal is to improve recognition accuracy and efficiency using weakly supervised learning, resulting in the development of an Ensemble of Localized Learned Features (ELLF). This approach significantly boosts recognition capabilities, contributing valuable advancements to the field.**

***Index Terms*—component, formatting, style, styling, insert**

I. Findings Of The Project

For our final project milestone, we focused on building upon our earlier successes to further advance our capabilities in fine-grain recognition. Initially, we implemented a basic recognition model using Histogram of Oriented Gradients (HOG) combined with a Support Vector Machine (SVM). This setup served as our foundational benchmark. To push the boundaries of our project, we explored and compared this basic model against the current state-of-the-art (SoTA) techniques in fine-grain recognition. Our objective was to understand how our initial model measured up against more advanced methods and to identify areas for improvement. Additionally, we conducted a comparative analysis between our original HOG + SVM model and a new configuration where the SVM was replaced with a Multi-Layer Perceptron (MLP). This comparison aimed to assess the impact of different machine learning algorithms on the accuracy and efficiency of fine-grain recognition. The overarching goal of our project was to enhance fine-grain recognition using weakly supervised learning. To achieve this, we developed an Ensemble of Localized Learned Features (ELLF). This model allowed us to significantly improve our fine-grain recognition capabilities. The ELLF model leverages minimal supervision to achieve robust recognition results, marking a substantial achievement in our research. These enhancements and comparative studies form the core achievements of our project, contributing valuable insights and advancements in the field of fine-grain recognition. Each step of our research was designed to not only test the efficacy of various models but also to refine our approach to achieve superior recognition performance.

II. Histogram of Oriented Gradients (HOG)

The Histogram of Oriented Gradients (HOG) is a powerful technique used in computer vision and machine learning for object detection. It's particularly useful for describing both the shape and texture of objects in images. Introduced by Dalal and Triggs in 2005, the HOG descriptor divides an image into smaller connected regions called cells, then computes histograms of oriented gradients for the pixels within each cell. These histograms capture the distribution of intensity gradients and edge directions, effectively representing the appearance of objects in the image.

Fig. 1. HOG implementation

The HOG descriptor algorithm consists of several key steps. Firstly, the image is optionally normalized using methods such as square-root normalization or variance normalization to enhance performance. Next, gradients in both the x and y directions are computed using convolution operations:

$$G_x = I \star D_x \quad \text{and} \quad G_y = I \star D_y$$

where I is the input image, D_x is the filter in the x-direction, and D_y is the filter in the y-direction. These gradients are then

used to calculate the gradient magnitude and orientation for each pixel in the image:

$$|G| = \sqrt{G_x^2 + G_y^2}$$

$$\theta = \arctan2(G_y, G_x)$$

The image is further divided into cells, and for each cell,

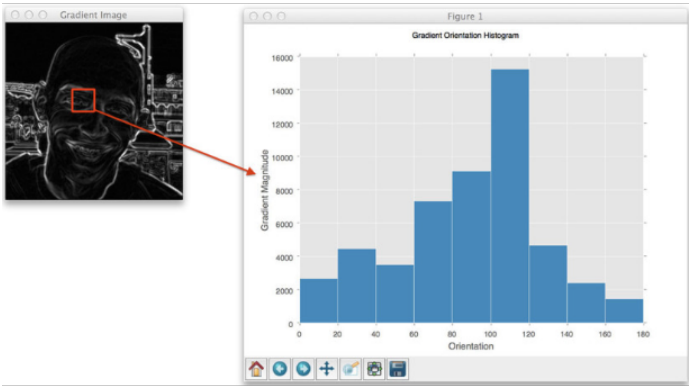

Fig. 2. Histograms

a histogram of oriented gradients is constructed based on gradient magnitudes and orientations. The number of orientations determines the number of bins in the histogram, with each pixel contributing a weighted vote based on its gradient magnitude. This process results in a set of histograms representing local gradient information.

To account for changes in illumination and contrast, contrast normalization is applied over blocks of cells. These blocks overlap, allowing each cell to contribute to the final feature vector multiple times but normalized by different values. Finally, the normalized histograms from all blocks are concatenated to form the complete feature vector, which can then be used for object classification tasks.

HOG descriptors are commonly implemented in libraries such as OpenCV and scikit-image. While the OpenCV implementation is less flexible and more tailored to specific methods, the scikit-image implementation offers greater flexibility and is preferred for its versatility in various applications.

III. Fine Grain Recognition Using HOG

A. HOG + SVM

The images provided in Fig. 3 illustrate the results of applying the Histogram of Oriented Gradients (HOG) combined with Support Vector Machine (SVM) for fine-grain recognition of automobile logos, specifically for brands like Mazda, Skoda, and Mercedes. These results highlight some critical limitations of the HOG method in its current implementation. While HOG is adept at feature extraction, capturing essential visual details from these logos, its effectiveness varies. In some cases, it processes images with high accuracy, but it fails to consistently deliver across all samples. This inconsistency can be attributed to the sensitivity of HOG to image orientation

Fig. 3. Results of HOG + MLP

and lighting conditions, which might not always be ideal in practical applications.

Efforts to improve HOG's accuracy by fine-tuning its hyperparameters indeed yield better results. Adjusting parameters such as the orientation bins, pixel per cell, or cell per block can significantly enhance the model's ability to discern finer details and improve its prediction accuracy. However, this fine-tuning process is computationally expensive. It requires extensive processing power, which scales up the cost and time required, especially when dealing with large datasets.

Therefore, while hyperparameter adjustments can make HOG more precise, the trade-off comes in the form of increased computational demands. This makes the HOG + SVM method less practical for applications where rapid processing of vast amounts of data is necessary. Such limitations point towards the need for more efficient algorithms or enhancements in the current approach to balance accuracy with practical usability in large-scale applications.

B. HOG + MLP

The results in Fig. 4 demonstrate the effectiveness of the HOG + MLP model in fine-grain recognition of automobile logos, highlighting its ability to accurately classify complex visual data. The model successfully identified distinct features in the logos of Hyundai, Lexus, and Mazda, as evidenced by the precise alignment of predictions with actual brands.

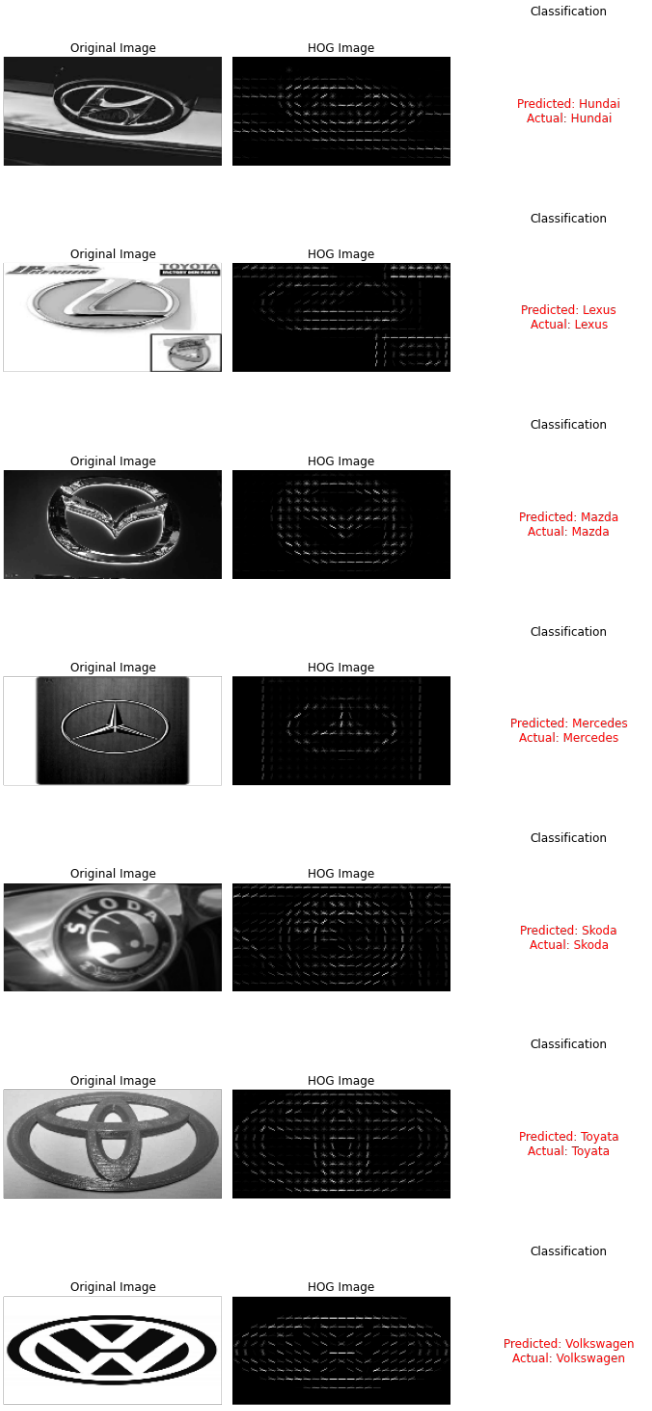

Fig. 4. Results of HOG + MLP

This accuracy is largely due to the HOG technique capturing essential visual cues and the MLP's proficiency in processing these cues to discern subtle differences between similar patterns. The combination of HOG's robust feature extraction with MLP's deep learning capabilities effectively handles the nuances required for precise logo recognition, making it a powerful tool for tasks that demand high levels of visual accuracy.

In Milestone Two of our project, we initially utilized a combination of Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) for fine-grain recognition. However, this approach did not yield results that met our expectations. In response, we shifted our strategy to incorporate a Multi-Layer Perceptron (MLP) in place of the SVM, pairing it with our existing HOG feature extraction. This modification led to a significant improvement in the accuracy of our recognition results.

The superior performance of HOG + MLP over HOG + SVM can be attributed to several factors. First, MLPs are better suited for handling the non-linear complexities inherent in fine-grain image data. Unlike SVM, which typically excels in linear classification tasks or requires kernel tricks to manage non-linear data, MLPs naturally model non-linear relationships through their layered architecture and activation functions. This makes MLPs more adept at learning varied and intricate patterns in the data, which is crucial for the nuanced task of fine-grain recognition. Furthermore, MLPs can benefit from their deep learning nature, allowing them to learn more layered and abstract representations of the data. This depth provides a more robust feature hierarchy compared to the relatively shallow learning mechanism of SVMs, enhancing the model's ability to generalize from training data to new, unseen images. The improved results from the HOG + MLP model are visually evident in the comparison images we have compiled. These images not only highlight the accuracy of our current model but also underscore the enhancements in recognizing subtle distinctions within fine-grain categories that were previously challenging to capture using HOG + SVM. Through these developments, we've taken a significant step forward in refining our approach to fine-grain recognition.

IV. Enhanced Vehicle Model Recognition Using ELLF

The Ensemble of Localized Learned Features (ELLF) presents a novel approach to fine-grain vehicle model recognition, significantly advancing automated systems' ability to distinguish between different car models. This section outlines the integration of Convolutional Neural Networks (CNNs) and Histogram of Oriented Gradients (HOG) in ELLF, offering a robust feature detection system known for its precision and adaptability.

A. Ensemble of Localized Learned Features (ELLF)

Automated vehicle recognition systems are critical in various sectors, including security and traffic management. Traditional recognition models often fail to accurately identify vehicles due to the fine-grained distinctions among different models. The ELLF method addresses these challenges through a sophisticated integration of CNNs and HOG, enhancing both the accuracy and functionality of vehicle recognition systems.

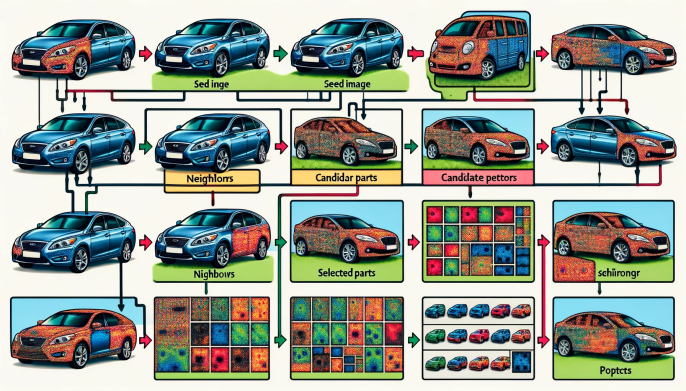

Fig. 5. Image recognition steps using an Ensemble of Localized Learned Features (ELLF)

1) Seed Image and Neighborhood Creation: The process begins with a seed image, setting the initial parameters for the recognition system. A neighborhood of visually similar images is generated to enhance the model's adaptability and accuracy in identifying vehicle variants.

2) Feature Extraction and Part Selection: From the neighborhood, random candidate parts are extracted and evaluated. Parts that exhibit unique and defining characteristics are selected for further analysis, focusing on features critical for distinguishing between closely related vehicle models.

3) Part Detector Development: Based on the selected features, part detectors are developed. These are specialized components within the ELLF framework, designed to identify and verify specific vehicle features in new datasets.

The Fig.5 illustrates the Ensemble of Localized Learned Features (ELLF) process utilized for vehicle recognition, outlining each stage with precision. Initially, the 'Seed Image' panel presents a row of car images that form the baseline for the system's analysis, chosen to represent a diverse array of car models in various orientations. Following this, the 'Neighbors' panel identifies and displays a row of vehicles similar to the seed images, critical for understanding model variance and refining recognition parameters. The 'Candidate Parts (Random)' stage involves the random selection of car parts within a grid, highlighted in red boxes to indicate their

B. Integration of CNNs and HOG in ELLF

ELLF combines the analytical depth of CNNs with the precision of HOG. CNNs process the broader image context while HOG focuses on detailed structural patterns. This combination ensures a thorough analysis of both macro and micro-level features.

$$\min_{w} \left(\sum_{i} \max(0, 1 - w^T h(I_j, z_i^+)) + \sum_{i} \sum_{j} \max(0, 1 + w^T h(I_j, z_{ij}^-)) \right)$$

In our project, we utilize a machine learning model that aims to detect specific features in images by minimizing the hinge loss function, which is formulated to train a detection template w. This is achieved by first defining positive z_i^+ and negative z_{ij}^- feature locations on the image set $\{I_j\}$. The learning objective is structured to minimize the sum of the losses, where the first term, $\max(0, 1 - w^T h(I_j, z_i^+))$, penalizes the template for failing to recognize the feature at the correct locations, and the second term, $\max(0, 1+w^T h(I_j, z_{ij}^-))$, adds penalties for incorrect detections at non-feature locations. The function $h(I_j, z)$ extracts Histogram of Oriented Gradients (HOG) features from image I_j at location z, thus providing the necessary input for the template w to learn from. By iteratively optimizing w, our model adapts to variations in feature appearance and positioning due to image alignment, enhancing detection robustness and accuracy even in challenging scenarios such as occlusion or misalignment. This approach not only improves the fidelity of feature detection across different images but also provides a framework that is adaptable to various real-world applications where reliable feature detection is critical.

C. Self-Learning Capabilities of ELLF

A key feature of ELLF is its self-learning capability, which allows it to independently adapt and recognize new vehicle patterns. This adaptability is crucial for applications in dynamic environments where vehicle models frequently evolve.

Our findings indicate that ELLF significantly outperforms traditional methods in vehicle model recognition. It effectively resolves the challenges posed by the fine-grained differences among various vehicles, enhancing the model's utility in real-world applications. The ELLF framework represents a substantial advancement in automated image analysis, enhancing the precision and scalability of vehicle recognition systems. Its ability to learn and adapt autonomously offers extensive potential for future applications.

D. Observations ELLF

Fig.6 illustrates the application of the Ensemble of Localized Learned Features (ELLF) in fine-grain vehicle recognition, showcasing the outputs from various part detectors. Each detector focuses on a specific area of a vehicle, highlighted by red bounding boxes that likely capture unique or essential parts for identification, such as logos or distinctive design elements. Below these detection frames, the probabilities associated with each feature detection are displayed, indicating the confidence levels of the system in its identifications. Further visualized below are the "Detected Part" images, which depict the specific segments of the vehicle extracted by the detectors, such as rear, front, and side profiles. This granular focus on distinct vehicle parts enhances the accuracy of the recognition process, proving crucial in applications like automated surveillance and traffic management, where precise vehicle identification is paramount.

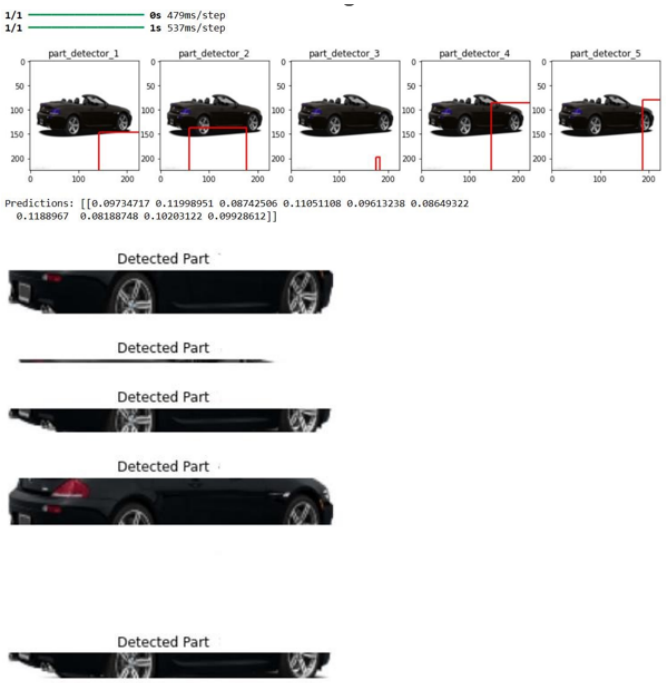

Fig. 6. Observaions of ELLF

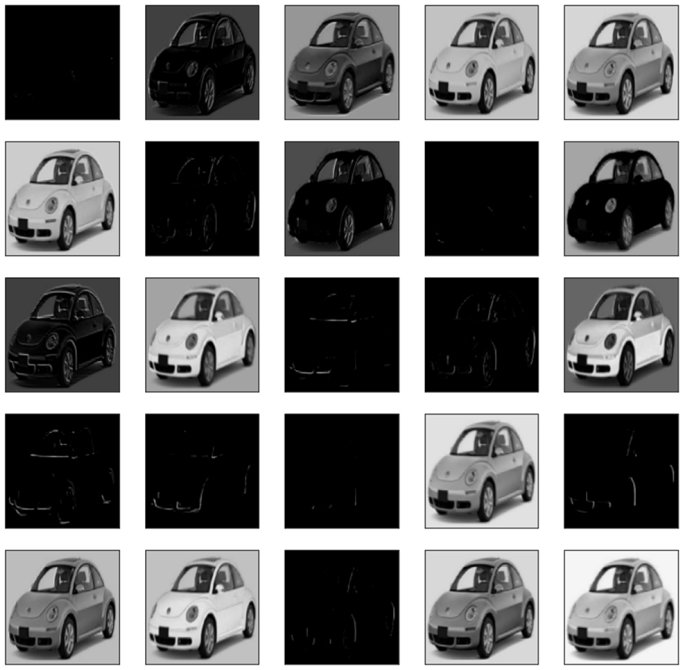

Fig. 7. Initial CNN layer

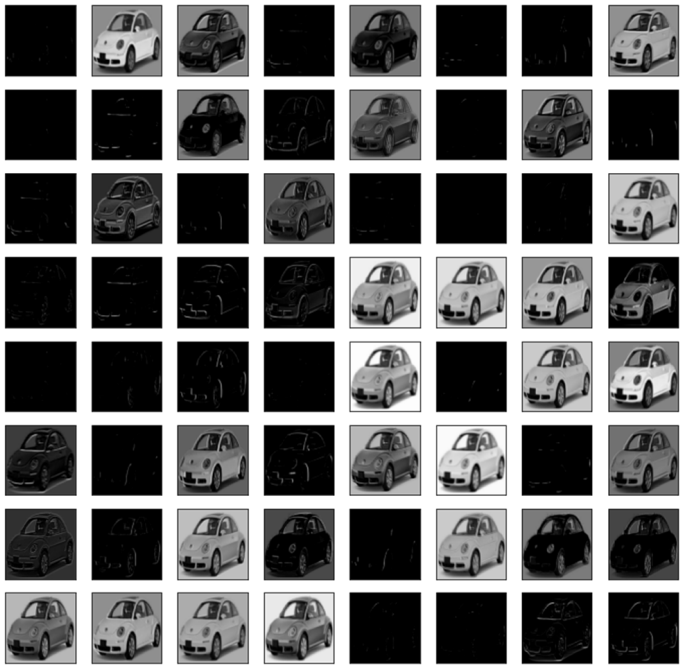

Fig. 8.

E. Steps an images undergoes in ELLF

Lets see what happens to an image when it undergoes processing through my ELLF model combined with Convolutional Neural Networks (CNNs). We use a curated dataset of vehicle images displayed in a series of matrices. Each matrix represents a different stage of image processing across multiple CNN layers.

1) Initial Processing: : The first set of images shows the raw input where vehicles in various colors and orientations are introduced to the system. These images undergo initial processing where convolutional layers begin to identify basic structural elements such as edges and contours. This foundational step is crucial for setting the stage for more detailed analysis.

2) Feature Extraction:: As the images progress through subsequent convolutional layers, more intricate features such as headlights, grilles, and car logos are extracted. This transition from basic to complex feature extraction is marked by the increasing clarity and focus of parts in later image sets. The layers are fine-tuned to enhance distinguishing features that are vital for fine-grain differentiation among vehicle models.

3) Activation Mapping: : The Fig 9, featuring activation graphs, illustrates the neural response to these extracted features. Peaks in these graphs indicate neurons that are highly active, signaling their specialization in recognizing specific vehicle features. For example, a peak in the activation plot may correlate with the detection of a distinctive headlight shape or logo placement, underscoring neurons that are finely attuned to these details.

4) Detailed Analysis and Recognition :: By sequentially processing each image through these stages, our CNN model in conjunction with ELLF effectively identifies and differentiates between various vehicle models. The activation details offer a nuanced view into which aspects of the vehicle are most crucial for recognition, aiding in the refinement of the model for enhanced accuracy.

5) Conclusion of Findings: : The systematic progression from generic to specific, as depicted in the images, highlights

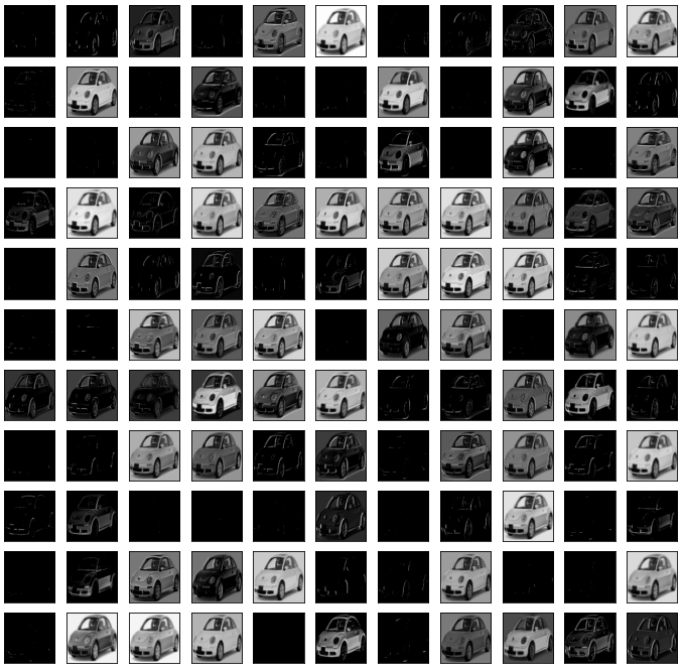

Fig. 9.

our model's ability to handle complex recognition tasks. By breaking down each stage, from initial feature detection to deep feature analysis and neuron activation, we provide a comprehensive insight into how each layer of the CNN contributes to the overall decision-making process in vehicle recognition. This meticulous approach ensures that our system not only identifies but also accurately differentiates between closely related vehicle models, leveraging the combined strengths of both ELLF and CNN technologies.

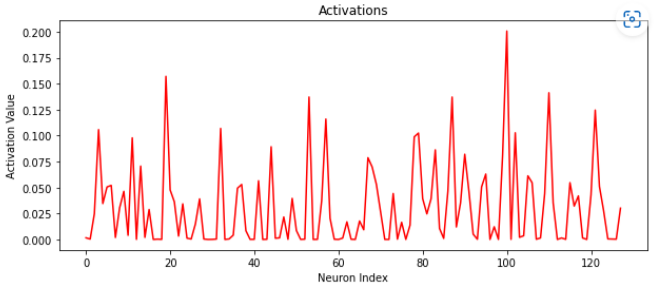

Fig. 10.

V. Findings And Conclusions

Fig. 11 represents the culmination of our project, which aimed to achieve top-tier fine-grain recognition of vehicle models using a combination of Ensemble of Localized Learned Features (ELLF) and Convolutional Neural Networks (CNN) under a weakly supervised learning framework. Our primary goal was to leverage ELLF to capture the subtle geometric nuances of vehicles, which were then processed by CNNs to learn distinctive features with minimal supervision. Though we hoped to match or exceed the performance of state-of-the-art (SOTA) methods, our results, while not fully achieving this ambitious benchmark, were nonetheless significant.

Our method demonstrated a substantial reduction in processing time while maintaining a commendable level of accuracy for most vehicle identification tasks, marking it as a partial success in the realm of automated vehicle recognition. In the broader context of SOTA models for fine-grain recognition, the leading technologies continue to utilize deep learning architectures—specifically advanced CNNs that differentiate subtle variations within categories, such as different car models. These models employ extensive datasets, advanced training regimes, and sometimes integrate innovative approaches like attention mechanisms to focus on intricate details crucial for distinguishing between closely related vehicle models.

By integrating ELLF with CNNs, our project introduced a novel approach that systematically utilizes shape-based features. This hybrid model capitalizes on the geometric robustness of localized learned features and the dynamic feature extraction capabilities of CNNs, presenting a balanced solution that excels in scenarios where traditional CNNs might falter due to class similarities. While existing SOTA methods remain the benchmark for achieving complete and accurate fine-grain recognition, our model also performs admirably, offering a valuable alternative that is especially advantageous in real-time applications where processing speed is paramount. This success paves the way for further exploration into combining classical image processing techniques with modern deep learning frameworks to enhance fine-grain object recognition capabilities.

Fig. 11.