Trustworthy Edge AI: Explainable MobileViT-S with Accurate Visual and Textual Reasoning

Shubham Limbachiya, Smit Patel

Abstract

Modern computer vision models such as CNNs and Vision Transformers achieve high accuracy but they also acts as black boxes. This lack of transparency becomes a serious issue when deploying models where decisions may directly affect users in many ways. MobileViT-S is a lightweight Vision Transformer designed for less resource environments. It provides strong accuracy but the original architecture offers no explainability, interpretability stability checks and no mechanism to understand predictions. For edge deployment the explanations must be fast, lightweight and easy for a non expert. As AI is growing everywhere there is an increasing need for models which are understandable and trustworthy. Our project fills this gap by extending it into a fully explainable vision system

Methodology

We enhanced a pretrained MobileViT-S model by fine tuning it on CIFAR-10 and for explainability Grad-CAM was integrated through an automatic feature layer selection mechanism to generate reliable heatmaps. To measure consistency of reasons we applied small perturbations such as rotations, brightness changes and noise and then compared the resulting Grad-CAM maps to original using Structural Similarity Index Measure(SSIM) which produce an Interpretability Stability Score (ISS). Alongside visual explanations the system generates brief natural language descriptions of models focus, structure sensitivity and color reliance resulting in an interpretable solution.

How Explanation Works?

After MobileViT-S produces logits for an image, Grad-CAM is used to identify which spatial regions contributed most to the model's predicted class. Then forward and backward hooks to the final feature layer of MobileViT-S captures both the activation maps A_k and the gradients for the target class C. The importance of each channel is computed using global average pooling of gradients, $\alpha_k = \frac{1}{H \cdot W} \sum_{i,j} \cdot \frac{\partial y^c}{\partial A_{k,ij}}$

and the class specific heatmap is constructed as $CAM = \text{ReLU}(\sum_k . \alpha_k A_k)$, which is then normalized to [0,1] and upsampled to input resolution. This heatmap highlights the regions that was used for classification.

- For reliability of explanation we generate an another Grad-CAM for a perturbed version of same image by using small rotations, brightness shifts and Gaussian noise and compare it with original heatmap. Both maps are min-max normalized by using,

$$CAM_{\text{norm}} = \frac{CAM - \min(CAM)}{\max(CAM) - \min(CAM) + 10^{-6}}$$

and their structural similarity is measured using the Structural Similarity Index (SSIM) metric which forms the Interpretability Stability Score,

$$ISS = SSIM(CAM_{\text{orig}}, CAM_{\text{pert}})$$

A higher ISS indicates that the model's reasoning is stable under small input changes which is very much needed to create trust.

Along with that our model also generates a short natural language explanation derived directly from the heatmap. It analyzes the region of maximum activation (like "upper-left", "center-right", etc.), the edge strength in the gradient area using Sobel filters and the most dominant color within the CAM mask. These values are filled in structured templates which makes the explanation to remain consistent in style but still showing the image specific visual explanation. As a result, each prediction includes a class label, probability, GradCAM heatmap, ISS score and a human readable description of models focus making it transparent and trustable.

Problem Definition

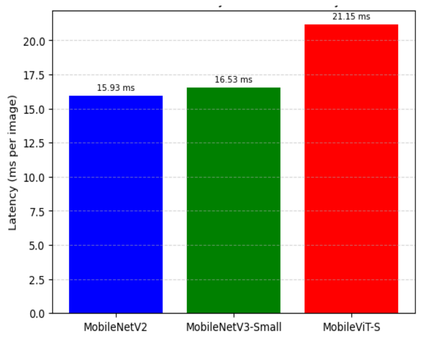
While compact object detection models are suitable for deployment there are key challenges like:

How can we guarantee the model's predictions are interpretable and trustworthy?

How do we evaluate the stability of explanations across variations in input images?

Traditional accuracy metrics fail to describe why the model predicted a class. The problem is that we have many pipelines with high accuracy but no sign of any reasons behind the prediction. This means that they work as a black box where the processing done for the results is hidden. Resulting predictions with no highlighted features and motive of result.

Inference and Accuracy



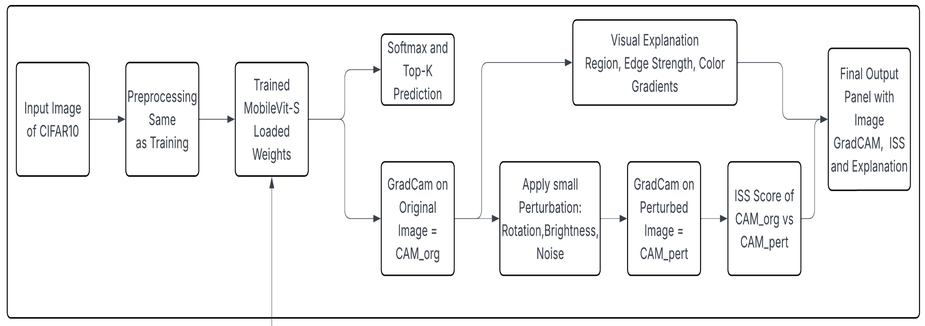
MobileViT-S has higher latency due to transformer based attention rollout and high Parameters as compared to other two models.

Models	Param. (in billions)	Memory (in Mb)	GFLOPs
MobileNet-V2	3.4	5.8	0.3
MobileNet-V3	4.5	8.5	0.06
MobileVit-s	5.6	18.86	1.6

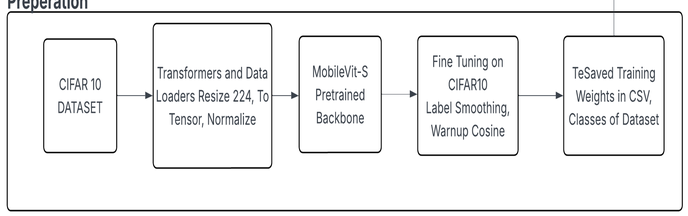
Comparison between each model in terms of their size and FLOPs in Billions, which shows how many calculations it can perform in one Second.

Working Of Whole Model

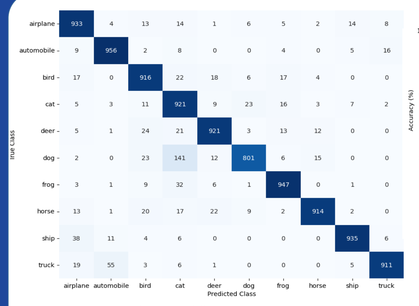
Explainability Pipeline



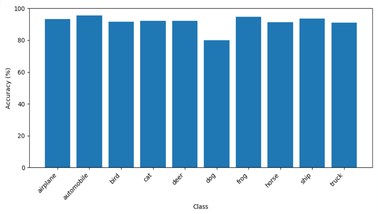
Training And Model Preperation



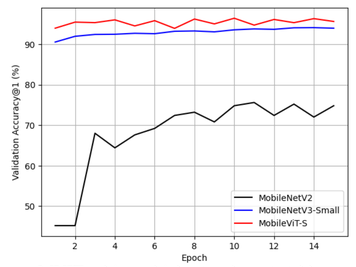
Results



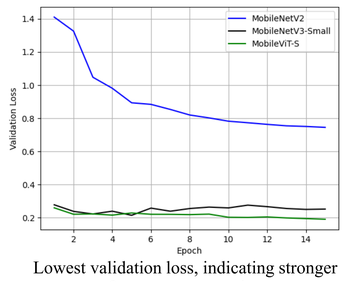
Accurately classifies most classes, but confuses most among Dog-Cat



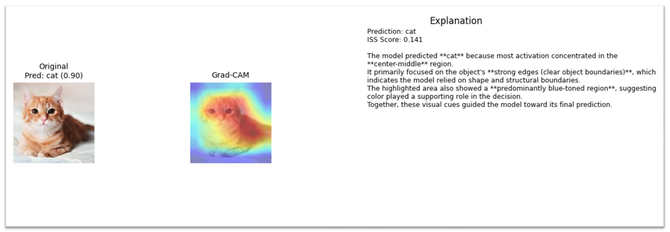
MobileViT-S Per Class Accuracy. It maintains high accuracy across most CIFAR-10 classes



MVIT achieves highest and most stable accuracy across epochs



Lowest validation loss, indicating stronger feature representation



Results with Explainability of what regions of the Image were responsible for Prediction.

Conclusion

Our results shows that MobileViT-S with Grad-CAM and ISS provides a deeper understanding of how the model makes decisions and how those decisions are changed under small changes irrespective of the classification accuracy. The heatmaps consistently highlighted the object showing that the model uses meaningful visual to show rather than just predicting as blackbox. Overall, our work demonstrates the importance of transparent and stable reasoning of an AI and shows how our pipeline directly solves the gap by providing interpretability of the model.

Future Work

To make it more reliable and useful model in future we will focus on improving the precision and stability of the explanations and ISS by expanding the pipeline to support a wider range of datasets and visual domains. An idea is to also apply this model to medical images where interpretability is very critical. By refining stability measures, enhancing region specific reasoning and adapting to sensitive real world use, we want to develop an explainable vision system that can be trusted by providing explanations more closer to judgement of an expert.