Hessian Sets: Uncovering Feature Interactions in Image Classification

Ayushi Mehrotra
Troy High School
am5715@rit.edu

Dipkamal Bhusal
Rochester Institute of Technology
db1702@g.rit.edu

Nidhi Rastogi
Rochester Institute of Technology
nxrvse@rit.edu

Abstract

Feature attribution methods explain model predictions by computing the contribution of individual features. However, these methods often overlook the impact of feature interactions, which play a crucial role in tasks like image classification. In this work, we introduce Hessian Sets, a technique that leverages the Hessian matrix to detect and attribute pairwise feature interactions in image classifiers. We adapt Integrated Directional Gradients (IDG) to assign importance to these feature interaction sets. By integrating segmentation masks from the Segment Anything Model (SAM), we provide more interpretable and concise explanations. Our initial experiments on the Imagenette dataset demonstrate that our method produces sparse, interpretable feature attributions while effectively capturing important interactions. This is a work in progress, and we present preliminary results to highlight the potential of our approach for improving explainability in image classifiers.

1 Introduction

Feature attribution methods, such as Integrated Gradients [10] and SHAP [7], have gained prominence in interpreting black-box models by quantifying the importance of individual features $x_i \in \mathbf{x}$. These techniques are essential for decision validation and bias detection in domains like healthcare, cybersecurity, and autonomous systems. However, they predominantly focus on the marginal effect of individual features, neglecting an equally crucial phenomenon: **feature interaction**. Feature interactions, where sets of features $\mathcal{I} \subset \mathbf{x}$ jointly influence the model's prediction, are particularly relevant in image classification tasks, where pixel and object dependencies often carry significant semantic information.

While feature interaction methods have seen progress in fields like natural language processing [9] and recommender systems [11], the application of these approaches to image classifiers remains underexplored. Existing concept-based methods attribute importance to image regions but often rely on segmentation or image-crops, limiting their ability to capture joint feature behaviors effectively [3] [2].

In this work, we address this gap by introducing **Hessian Sets**, an approach that leverages the Hessian matrix to detect pairwise feature interactions within images. Our method recursively merges these interactions into larger sets, offering a robust measure of feature interactions while maintaining interpretability. Additionally, we adapt **Integrated Directional Gradients (IDG)** to attribute importance to interaction sets. By integrating segmentation masks from the Segment Anything Model (SAM), we ensure that feature interactions are tied to meaningful image regions, leading to more concise and interpretable heatmaps.

2nd Workshop on Attributing Model Behavior at Scale, (NeurIPS 2024).

2 Feature Interaction Detection

In order to find the set of feature interaction sets $\mathcal{I}_i \in \mathcal{S}$ in a high-dimensional feature space, we draw inspiration from gradient-based attribution methods.

2.1 Hessian Matrix

We use the Hessian matrix to detect pairwise feature interactions. However, due to gradient saturation [10], we do not rely on the Hessian for attributing feature interactions. Since higher-order interactions imply that any subset of those features will also interact, we recursively identify pairwise interactions and combine them into an interaction set, $\mathcal{I}$, continuing this process until no further interactions are found that meet the threshold μ or the maximum number of elements allowed in a feature interaction set ν.

$$\mathbf{H}_f = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

The Hessian requires a point x_j to calculate pairwise interactions with all features $x_i \in \mathbf{x}$. In the current work, we propose to fix x_j as the highest attributed feature by Integrated Gradients in a segmentation mask produced by Segment Anything Model (SAM) [6]. In reality, this point may be chosen anywhere on the image, allowing for human-guided feature interaction sets as well.

Algorithm 1 Hessian Set Algorithm

function GENERATESETS(example $\mathbf{x}$, model f, gradient $\nabla f(\mathbf{x})$, index j, threshold μ, max elements ν)
  $\mathbf{H}_j \leftarrow \frac{\partial}{\partial x_i}(\nabla f(\mathbf{x}))_j$ ▷ Create Hessian
  $\mathbf{H}'_j \leftarrow \arg\min_i (\mathbf{H}_j[i] > \mu)$ ▷ Consider interactions above threshold
  next $\leftarrow \{\}$
  for all $n \in \mathbf{H}'_j$ **do**
    if $n = 0$ **then**
      continue
    else if $n \notin \mathcal{I}$ and $|\mathcal{I}| < \nu$ **then**
      $\mathcal{I} \leftarrow \mathcal{I} \cup n$ ▷ Add interaction to current feature interaction set
      next $\leftarrow$ next $\cup\, n$
    end if
  end for
  for all $m \in$ next **do**
    GENERATESETS($\mathbf{x}$, f, $\nabla f(\mathbf{x})$, argmin$_j\{\text{next}[j] = m\}$, μ) ▷ Find more interactions
  end for
  return $\mathcal{I}$
end function

3 Feature Interaction Attribution

In order to attribute each feature interaction set $\mathcal{I}$, we draw upon Integrated Directional Gradients (IDG) [9], originally proposed for natural language processing.

3.1 Integrated Directional Gradients

Given a feature interaction set $\mathcal{I}$, baseline $\mathbf{b}$, and image $\mathbf{x}$, we calculate the importance based on the constructed value function for $\mathcal{I}$ using Equations 1 to 6.

$$a_i = \begin{cases} x_i - b_i & \text{if } x_i \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\hat{a} = \frac{a}{||a||} \tag{2}$$

$$\nabla_{\mathcal{I}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \hat{a} \tag{3}$$

$$\text{IDG}(\mathcal{I}) = \int_{\alpha=0}^{1} \nabla_{\mathcal{I}} f(\mathbf{b} + \alpha(\mathbf{x} - \mathbf{b}))d\alpha \tag{4}$$

$$d(\mathcal{I}) = ||\text{IDG}(\mathcal{I})|| \tag{5}$$

$$v(\mathcal{I}) = \sum_{T \subseteq \mathcal{I}} d(T) \tag{6}$$

Equation (1) generates the feature difference vector $\mathbf{a}$ based on the feature interaction set $\mathcal{I}$. The following equation constructs the directional vector for $\mathbf{a}$. Equation (3) computes the directional gradient, and Equation (4) defines the Integrated Directional Gradient (IDG) algorithm. Equation (5) defines the Harsanyi dividend [4] for a group of features. Finally, Equation (6) calculates the importance of a feature interaction set. Since computing the Integrated Directional Gradient for every possible subset is impractical in high-dimensional feature spaces, we use random sampling in Equation (6).

Since Equation (4) includes an integral, we approximate it with a Riemann's sum.

$$\text{AIDG}(\mathcal{I}) = \frac{1}{m+1} \sum_{k=0}^{m} \nabla_{\mathcal{I}} f(\mathbf{b} + \frac{k}{m}(\mathbf{x} - \mathbf{b})) \tag{7}$$

4 Experiments

Table 1: Sparsity evaluation for proposed method and Integrated Gradients

Method	Sparsity
Our method	0.9565 ± 0.016
Integrated Gradient	0.8235 ± 0.407

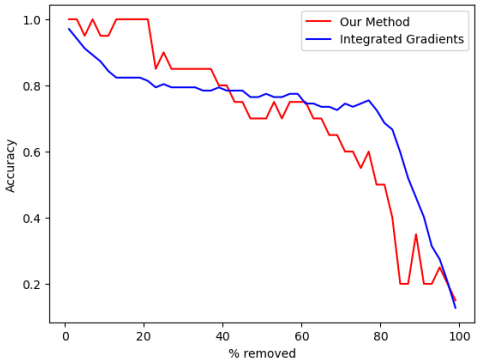

Figure 1: Evaluation results in MoRF using ROAD

We evaluate our methods on the Imagenette dataset using a ResNet50 model [5]. To measure the effectiveness of our approach, we use two metrics: sparsity [1],which measures the conciseness of explanations and ROAD (RemOve And Debias) [8], which assess how faithfully the feature interaction attribution method reflects the model's behavior. We set the threshold μ to half of x_j attribution value and cap the maximum number of features ν in a feature interaction set at 2000. We then compare our method's performance against the Integrated Gradient [10].

Table 1 shows that our proposed approach generates sparse explanations, which is one of the desirable characteristics of feature attribution method. Sparser explanations only attribute highly relevant features resulting in more comprehensible explanations [1].

We adopt the MoRF (Most Relevant First) removal strategy for ROAD analysis. Figure 1 illustrates the evaluation results on Imagenette. In the MoRF strategy, an attribution method with a sharper drop in accuracy as k most important features are removed indicate higher faithfulness. Initially, we observe that Integrated Gradients shows a steeper drop compared to our proposed method, which could be attributed to our method assigning similar importance to a broader set of features. However, as more features are removed, our method appears to have captured a larger set of discriminative features.

We illustrate some representative figures of heatmaps in Figure 2 for comparison.

Figure 2: Example instances of our methodology on an ImageNette image. Each mask is attributed with a single value that represents the importance of the feature interaction set.

5 Conclusion

In this work, we propose an approach to capture feature interactions in image classifiers through Hessian Sets and Integrated Directional Gradients. Our method successfully identifies and attributes joint influences of features, offering a more comprehensive understanding of model predictions compared to traditional feature attribution methods. Initial results on the Imagenette dataset show that our method generates sparser and more interpretable explanations while maintaining faithfulness to the model's behavior. Although these findings are promising, this work is still in progress. Future steps include fine-tuning the interaction set construction and further evaluating the method on more diverse datasets and models with various attribution evaluation metrics.

References

[1] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pages 1383–1391. PMLR, 2020.

[2] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for ex-

plainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023.

[3] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[4] John C Harsanyi and John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *Papers in game theory*, pages 44–70, 1982.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[7] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[8] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022.

[9] Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878, 2021.

[10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[11] Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. *Advances in neural information processing systems*, 33:6147–6159, 2020.