Towards Fine-Grained Interpretability: Counterfactual Explanations for Misclassification with Saliency Partition

Supplementary Material

6. Algorithm

The counterfactual generation procedure is depicted in Algorithm 1. In main text and in the algorithm, some superscripts and subscripts are omitted for simplicity. $h^{(0)}$ and h both refer to the original feature map after feature extraction.

7. Performance evaluation

7.1. Evaluation on the Compact Activation Score

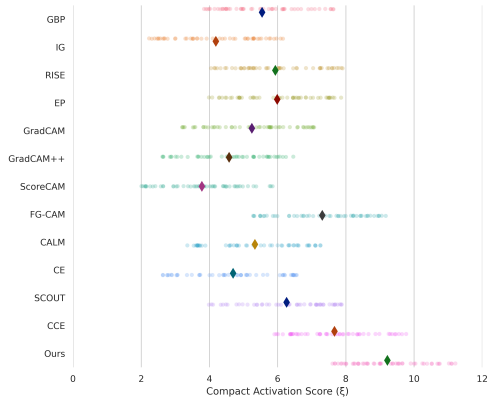

Figure 8. A comparative evaluation of the performance of multiple methods in generating 50 attributive explanation based on the Compact Activation Score (ξ) metric.

As shown in Fig. 8, we randomly selected explanations generated for 50 misclassified samples using different attribution methods and evaluated them on a fine-grained metric, ξ. The results show the distribution of ξ scores for the 50 explanations generated by each method. For our generated explanations, the scores were primarily concentrated within the range of 8 to 12, which corresponds to the highest scores.

7.2. Evaluating the interpretability of comparative explanation methods

To evaluate the quality of explanations generated by multiple comparative methods, we conducted assessments across five different aspects for fine-grained level: Locality (**Loc.**), Semantic relevance (**Sem.**), Stability (**Stab.**), Log of Odds (**LOdds**), and Area Under the Curve (**AUC**). As shown in Fig. 9, our method demonstrates excellent performance across five metrics in fine-grained level comparative explanations.

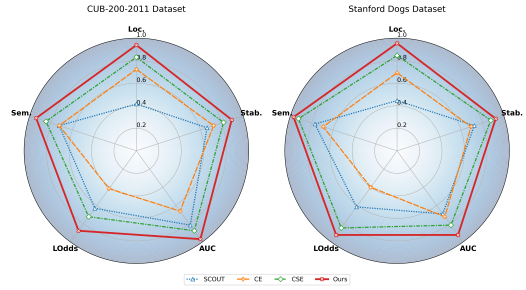

Figure 9. Performance evaluation results of multiple explanation methods across five aspects.

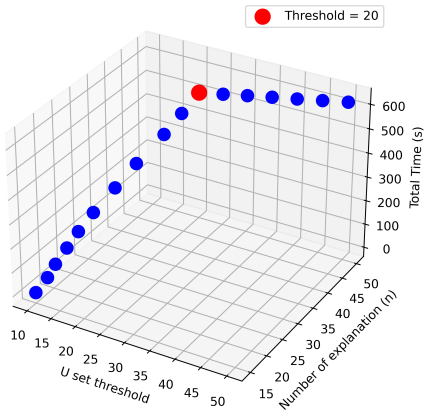

Figure 10. The impact of U set threshold on the number of explanations generated and the total time consumed.

8. Ablation Study

8.1. Threshold for the number of selected elements in set U

To explore the impact of selecting the threshold for the number of feature units in U set on the overall framework's performance, we analyze the relationship between each threshold and the success rate of generating explanations, as well

Algorithm 1: Counterfactual generation procedure of the proposed FG-VCE framework.

Input: A misclassified sample I and a correctly classified sample set $U = \{u_1, u_2, \cdots\}$, both from class a. A pre-trained feature extractor f with a classifier f_c. A set of of spatially localized kernels $\tilde{G}$.

Output: The feature map h^* after the change of model's prediction and the corresponding Shapley value map s^*.

Extract features using pre-trained model f: $h^{(0)}, h_U = f(I), f(U)$;

/* Saliency Partition */

Compute model's original prediction p_h for class a and the contribution matrix $p_{\tilde{H}}$:
$p_{h^{(0)}} = y_a \log f_c(h^{(0)}),\ p_{\tilde{H}^{(0)}_{[k,\,:,\,:]}} = y_a \log f_c(h^{(0)} \odot \tilde{G}_{[k,\,:,\,:]})\ \forall \tilde{G}_{[k,\,:,\,:]} \in \tilde{G}$;

Compute initial Shapley value map $s^{(0)}$ with Eq. 1: $s^{(0)} = p_h \cdot \mathbf{1} - p_{\tilde{H}}$;

Obtain feature $h^{(0)}_{s^*}$ with highest Shapley value $s^* = \max s^{(0)}$; // s^* is the max value in current iteration

foreach h_u *in* h_U **do**

  $p_h = y_a \log f_c(h_u),\ p_{\tilde{H}_{[k,\,:,\,:]}} = y_a \log f_c(h_u \odot \tilde{G}_{[k,\,:,\,:]})\ \forall \tilde{G}_{[k,\,:,\,:]} \in \tilde{G}$;

  $\tilde{s}_u = \mathrm{Top}_m(p_h \cdot \mathbf{1} - p_{\tilde{H}})$;

end

/* Fine-Grained Counterfactual Generation */

for *t=1..max_iteration* **do**

  Search for $[k, i, j]^{(t)}$ that satisfies overall objective function in Eq. 9;

  Obtain $h^{(t)}$ by replacing $h^{(t-1)}_{s^*}$ with $h_{\tilde{s}_{[k,i,j]^{(t)}}}$;

  if $\arg\max f_c(h^{(t)}) == c$ **then**

    $h^* \leftarrow h^{(t)}$, $s^* \leftarrow s^{(t)}$; // s^* is re-assigned to be the Shapley value map of h^* at last

    break;

  else

    Compute $p_{h^{(t)}} = y_a \log f_c(h^{(t)}),\ p_{\tilde{H}^{(t)}_{[k,\,:,\,:]}} = y_a \log f_c(h^{(t)} \odot \tilde{G}_{[k,\,:,\,:]})\ \forall \tilde{G}_{[k,\,:,\,:]} \in \tilde{G}$;

    Compute Shapley value map $s^{(t)}$ in iteration t with Eq. 1: $s^{(t)} = p_{h^{(t)}} \cdot \mathbf{1} - p_{\tilde{H}^{(t)}}$;

    Obtain feature $h^{(t)}_{s^*}$ with highest Shapley value $s^* = \max s^{(t)}$;

  end

end

as the total time consumed for 50 misclassified samples. According to Fig. 10, when the threshold is set to 20, explanations for all 50 misclassified samples are successfully generated. Beyond 20, the number of generated explanations does not change, but the total time consumed gradually increases due to the increasing computational complexity. Therefore, we select 20 as the optimal U threshold setting.

8.2. Impact of Gaussian kernel (σ) on the Shapley map

To determine the optimal value of σ, we adjusted different σ values within the SP module to observe their impact on the generated explanations in terms of the metrics Insertions (Ins.), Deletions (Del.), and the Compact Activation Score (ξ). As shown in Fig. 11, when the sigma value is set to 0.8, the combined evaluation of the three metrics reaches the optimal value. Therefore, we selected 0.8 as the optimal σ value in our method.

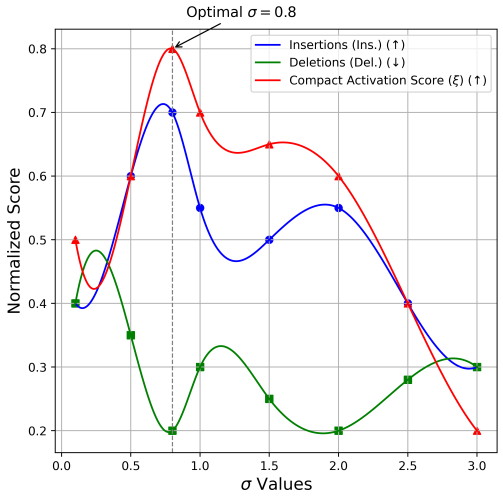

Figure 11. The performance impact curve of different σ values on the explanations generated by our method.

8.3. Impact of hyperparameter t

We conducted additional validation on the hyperparameter t, as shown in Figure 12.

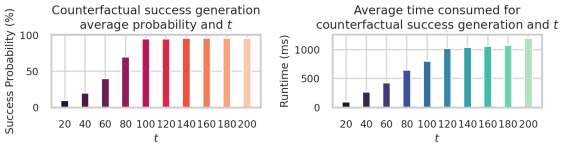

Figure 12. The impact of t on counterfactual generation. When t =100, the time consumed is minimal when the probability of counterfactual generation nearly reaches its peak.