Supplementary Materials for Adaptive Explanations via Baseline Exploration-Exploitation

1 EVALUATION METRICS

There is no single measure or test set which is generally acceptable for evaluating explanation maps. Hence, in order to ensure comparability, the evaluations in this research follow earlier works [3, 5, 10, 14]. In general, the various tests entail different types of masking of the original input according to the explanation maps and investigating the change in the model's prediction for the masked input compared to its original prediction based on the unmasked input. There are two variants for these tests which differ based on the class of reference. In one variant, the difference in predictions refers to the ground-truth class, and in the second variant, the difference in predictions refers to the model's original top-predicted class. In the manuscript, we report results for both variants and dub the first variant as 'target' and the second variant as 'predicted', respectively.

In what follows, we list and define the different evaluation measures used in this research:

(1) Average Drop Percentage **(ADP)** [3]: ADP = $100\% \cdot \frac{1}{N}\sum_{i=1}^{N}\frac{\max(0, Y_i^c - O_i^c)}{Y_i^c}$, where N is the total number of images in the evaluated dataset, Y_i^c is the model's output score (confidence) for class c w.r.t. the original image i. O_i^c is the same model's score, this time w.r.t. to a masked version of the original image (produced by the Hadamard product of the original image with the explanation map). The **lower** the ADP the better the result.

(2) Percentage of Increase in Confidence **(PIC)** [3]: PIC = $100\% \cdot \frac{1}{N}\sum_{i=1}^{N} 1(Y_i^c < O_i^c)$. PIC reports the percentage of cases in which the model's output scores increase as a result of the replacement of the original image with the masked version based on the explanation map. The explanation map is expected to mask the background and help the model to focus on the original image. Hence, the **higher** the PIC the better the result.

(3) Perturbation tests entail a stepwise process in which pixels in the original image are gradually masked out according to their relevance score obtained from the explanation map [5]. At each step, an additional 10% of the pixels are removed and the original image is gradually blacked out. The performance of the explanation model is assessed by measuring the area under the curve (AUC) with respect to the model's prediction on the masked image compared to its prediction with respect to the original (unmasked) image. We consider two types of masking:

   (a) Positive perturbation (**POS**), in which we mask the pixels in decreasing order, from the highest relevance to the lowest, and expect to see a steep decrease in performance, indicating that the masked pixels are important to the classification score. Hence, for the POS perturbation test, lower values indicate better performance.

   (b) Negative perturbation (**NEG**), in which we mask the pixels in increasing order, from lowest to highest. A good explanation would maintain the accuracy of the model while removing pixels that are not related to the class of interest. Hence, for the NEG perturbation test, lower values indicate better performance.

   In both positive and negative perturbations, we measure the area-under-the-curve (AUC), for erasing between 10%-90% of the pixels. As explained above, results are reported with respect to the 'predicted' or the 'target' (ground-truth) class.

(4) The deletion and insertion metrics [14] are described as follows:

   (a) The deletion (**DEL**) metric measures a decrease in the probability of the class of interest as more and more important pixels are removed, where the importance of each pixel is obtained from the generated explanation map. A sharp drop and thus a low area under the probability curve (as a function of the fraction of removed pixels) means a good explanation.

   (b) In contrast, the insertion (**INS**) metric measures the increase in probability as more and more pixels are revealed, with higher AUC indicative of a better explanation.

   Note that there are several ways in which pixels can be removed from an image [6]. In this work, we remove pixels by setting their value to zero. Gradual removal or introduction of pixels is performed in steps of 0.1 i.e., remove or introduce 10% of the pixels on each step).

(5) The Accuracy Information Curve (**AIC**) and the Softmax Information Curve (**SIC**) [10] metrics are both similar in spirit to the receiver operating characteristics (ROC). These measures are inspired by the Bokeh effect in photography [12], which consists of focusing on objects of interest while keeping the rest of the image blurred. In a similar fashion, we start with a completely blurred image and gradually sharpen the image areas that are deemed important by a given explanation method. Gradually sharpening the image areas increases the information content of the image. We then compare the explanation methods by measuring the approximate image entropy (e.g., compressed image size) and the model's performance (e.g., model accuracy).

   (a) The AIC metric measures the accuracy of a model as a function of the amount of information provided to the explanation method. AIC is defined as the AUC of the accuracy vs. information plot. The information provided to the method is quantified by the fraction of input features that are considered during the explanation process.

   (b) The SIC metric measures the information content of the output of a softmax classifier as a function of the amount of information provided to the explanation method. SIC is defined as the AUC of the entropy vs. information plot. The entropy of the softmax output is a measure of the uncertainty or randomness of the classifier's predictions. The information provided to the method is quantified by the fraction of input features that are considered during the explanation process.

2 EXPLANATION METHODS

(1) Grad-CAM (**GC**) [15] integrates the activation maps from the last convolutional layer in the CNN by employing global average pooling on the gradients and utilizing them as weights for the feature map channels.

(2) Grad-CAM++ (**GC++**) [3] is an advanced variant of Grad-CAM that utilizes a weighted average of the pixel-wise gradients to generate the activation map weights.

(3) Iterated Integrated Attributions (**IIA**) [2] an explanation approach that generalizes Integrated Gradients to an iterated integral.

(4) Integrated Gradients (**IG**) [18] integrates over the interpolated image gradients.

(5) Blur IG (**BIG**) [19] is concerned with the introduction of information using a baseline and opts to use a path that progressively removes Gaussian blur from the attributed image.
(6) Guided IG (**GIG**) [11] improves upon Integrated Gradients by introducing the idea of an adaptive path method. By calculating integration along a different path than Integrated Gradients, high gradient areas are avoided which often leads to an overall reduction in irrelevant attributions.
(7) LIFT-CAM (**LIFT**) [9] employs the DeepLIFT [16] technique to estimate the activation maps SHAP values [13] and then combine them with the activation maps to produce the explanation map.
(8) The FullGrad (**FG**) method [17] provides a complete modeling approach of the gradient by also taking the gradient with respect to the bias term, and not just with respect to the input.
(9) LayerCAM (**LC**) [8] utilizes both gradients and activations, but instead of using the Grad-CAM approach and applying pooling on the gradients, it treats the gradients as weights for the activations by assigning each location in the activations with an appropriate gradient location. The explanation map is computed with a location-wise product of the positive gradients (after ReLU) with the activations, and the map is then summed w.r.t. the activation channel, with a ReLU applied to the result.
(10) Ablation-CAM (**AC**) [7] is an approach that only uses the channels of the activations. It takes each activation channel, masks it from the final map by zeroing out all locations of this channel in the explanation map produced by all the channels, computes the score on the masked explanation map (the map without the specific channel), and this score is used to assign an importance weight for every channel. At last, a weighted sum of the channels produces the final explanation map.
(11) The Transformer attribution (**T-ATTR**) [5] method computes the importance of each input token by analyzing the attention weights assigned to it during self-attention. Specifically, it computes the relevance score of each token as the sum of its attention weights across all layers of the Transformer. The intuition behind this approach is that tokens that receive more attention across different layers are likely more important for the final prediction. To obtain a more interpretable and localized visualization of the importance scores, the authors also propose a variant of the method called Layer-wise Relevance Propagation (LRP), which recursively distributes the relevance scores back to the input tokens based on their contribution to the intermediate representations.
(12) Generic Attention Explainability (**GAE**) [4] is a generalization of T-Attr for explaining Bi-Modal transformers.

3 GRADIENT ROLLOUT IMPLEMENTATION

The Gradient Rollout (**GR**) technique is a modified version of the Attention Rollout (**AR**) [1] method, which differentiates itself by including a Hadamard product between each attention map and its gradients in the computation, rather than relying solely on the attention map. The GR method can be expressed mathematically as follows:

$$A'_b = I + E_h(A_b \circ G_b), \tag{1}$$

$$GR = A'_1 \cdot A'_2 \cdots A'_B. \tag{2}$$

where A_b is a 3D tensor consisting of the 2D attention maps produced by each attention head in the transformer block b, G_b is the gradients w.r.t. A_b. I is the identity matrix, B is the number of transformer blocks in the model, E_h is the mean reduction operation (taken across the attention heads dimension), and $\circ$ and $\cdot$ are the Hadamard product and matrix multiplication operators, respectively. Following this, GR proceeds with the original Rollout computation [1], resulting in the first row of the derived matrix (associated with the [CLS] token). Finally, this output is processed by truncating its initial element and reshaping it into a 14 × 14 matrix. The exact implementation of GR appears in our GitHub repository.

REFERENCES

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Oren Barkan, Yehonatan Elisha, Yuval Asher, Amit Eshel, and Noam Koenigstein. Visual explanations via iterated integrated attributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2073–2084, October 2023.

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[4] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

[5] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[6] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6970–6979, 2017.

[7] Saurabh Satish Desai and H. G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 972–980, 2020.

[8] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[9] Hyungsik Jung and Youngrock Oh. Towards better explanations of class activation mapping. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1316–1324, 2021.

[10] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957, 2019.

[11] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021.

[12] D Liu, R Nicolescu, and R Klette. Stereo-based bokeh effects for photography. *Machine Vision and Applications*, pages 1–13, May 2016.

[13] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017.

[14] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[16] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.

[17] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *NeurIPS*, 2019.

[18] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 3319–3328, 2017.

[19] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020.