Visual Explanation by Unifying Adversarial Generation and Feature Importance Attributions

Martin Charachon[1,2], Paul-Henry Cournède[2], Céline Hudelot[2], and Roberto Ardon[1]

[1] Incepto Medical, France
[2] MICS, Université Paris-Saclay, CentraleSupélec, France

Abstract. Explaining the decisions of deep learning models is critical for their adoption in medical practice. In this work, we propose to unify existing adversarial explanation methods and path-based feature importance attribution approaches. We consider a path between the input image and a generated adversary and associate a weight depending on the model output variations along this path. We validate our attribution methods on two medical classification tasks. We demonstrate significant improvement compared to state-of-the-art methods in both feature importance attribution and localization performance.

Keywords: Explainable AI · Deep learning · Classification · GANs · Medical image.

1 Introduction

While deep learning models are nowadays commonly used in the medical domain [1,8,14], a major limitation to their general adoption in daily clinical routines is the lack of explanations with respect to their predictions [16]. In this work, we focus on visual explanation methods [9,21,30]. They provide an additional image where regions of higher importance -for the model prediction- are expected to correlate with pathology location when the model prediction is correct or fail to otherwise. Thus, visual explanation images help clinicians to assign a level of confidence to a model.

Several contributions have been made [4,7,20,23,28] based on the generation of adversaries (images closely related to input images but that have a different model prediction), where visual explanation is defined as the difference between this adversary and the input image, or its regularized version [4,23]. They perform particularly well in highlighting global relevant regions for classification models that match expert expectations e.g. localized pathology. Despite efforts to produce visual explanations only containing relevant information to the model, residual noise still remains. On the other hand, path-based methods [17], which derive visual explanation from pixel-wise derivatives of the model, are built to detect pixel regions with high impact on the model prediction. They generally result in very noisy outputs. The main contribution of this paper is to unify both adversarial and path-based feature attribution methods (section 3). In the spirit of [26], we go a step further and follow a path between the input image and its adversary. We associate to each element of this path a weight reflecting feature importance through the model gradient. Visual explanation is then defined as the sum of all contributions along the path. We validated our method (section 4) on two classification tasks and publicly available data sets: slice classification for brain tumors localization in MRI, and pneumonia detection on chest X-Rays.

2 Related Works

Within the numerous contributions in visual explanation for classification models [5,9,10,19,21], our work is at the crossroads of two major families, gradient attribution methods and adversarial methods. In the following, let f be the classifier to explain, $\mathbf{x}$ the input image to which the classifier f is applied and $\mathcal{E}$ the visual explanation image.

Gradient attribution methods [21,24,25,26,29] generate visual explanations by associating an importance value with each pixel of $\mathbf{x}$ using back propagation of the classifier's gradient: $\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}) = \left\{\frac{\partial f}{\partial \mathbf{x}_i}(\mathbf{x})\right\}_i$ where index i enumerates all pixels of $\mathbf{x}$. In particular, [26] introduced the idea of integrating this gradient along a path of images to define the explanation as $\mathcal{E} = (\mathbf{x}-\mathbf{z}) \int_0^1 \frac{\partial f(\mathbf{z}+\lambda(\mathbf{x}-\mathbf{z}))}{\partial \mathbf{x}} \, d\lambda$ (where $\mathbf{z}$ is the null image or random noise). These methods generally produce visual explanations that contain relevant regions but are often corrupted by noise and have limited quality (see table 1).

In a different perspective, adversarial methods [3,4,7,23,28] propose to generate visual explanation by comparing the input image with a "close" generated adversarial example $\mathbf{x_a}$ and defining visual explanation by

$$\mathcal{E}(\mathbf{x}) = |\mathbf{x} - \mathbf{x_a}|. \tag{1}$$

The prediction of classifier f for the adversary $\mathbf{x_a}$ is expected to be different from that of the original image (e.g $f(\mathbf{x_a}) = 1 - f(\mathbf{x})$ for a binary classifier) and only differ from the input image on regions that are critical for the decision of f. Moreover, [3,23] advocate that $\mathbf{x_a}$ should belong to the distribution of real images (for instance by leveraging domain translation techniques [31]) in order to explain the behavior of f within the distribution of images it is expected to work on. Several works [4,18,23] also use the notion of a "stable" image $\mathbf{x_s}$ defined as the closest element to $\mathbf{x}$ (in the sense of norm $L_{1,2}$) generated by a comparable generation process as for $\mathbf{x_a}$ but classified by f like $\mathbf{x}$ ($f(\mathbf{x}) = f(\mathbf{x_s})$). The goal is to reduce reconstruction errors due to the generation process which are irrelevant to the visual explanation, then defined as $\mathcal{E} = |\mathbf{x_s} - \mathbf{x_a}|$. These methods have high performances in localizing pathological regions when acting on a classifier f trained to detect if there is any pathology (table 1).

3 Method

Compared to gradient attribution methods, adversarial approaches generate outputs that are smoother and more localized (table 1). However, no adversarial approach explicitly enforces that visual explanation values translate into importance values for f at the pixel level or at any higher scale. For instance, suppose $\mathbf{x}$ is a CT-scan and classifier f is influenced by regions containing bone tissues (which have high intensity in CT-scans). These regions should then be attenuated in the adversary $\mathbf{x_a}$ and appear with high intensity in the difference $\mathcal{E} = |\mathbf{x} - \mathbf{x_a}|$. This high intensity may not be directly related to the relative importance of bone regions for f but only result from their original intensity in $\mathbf{x}$. This case would poorly perform using the AOPC metric introduced in [17]. Experimentally, even when using $\mathbf{x_s}$, it is sometimes impossible to remove all irrelevant regions for f.

3.1 Combining gradient attribution and adversarial methods

Consider an image $\mathbf{x}$ or its "stable" generation $\mathbf{x_s}$ (depending on the chosen adversarial method) and its generated adversary $\mathbf{x_a}$. Following [23,26] we consider a differential path γ mapping elements $\lambda \in [0, 1]$ to the space of real images ([4,23]) and satisfying $\boldsymbol{\gamma}(0) = \mathbf{x}$ and $\boldsymbol{\gamma}(1) = \mathbf{x_a}$. From equation (1) we have

$$\mathcal{E}(\mathbf{x}) = |\mathbf{x_a} - \mathbf{x}| = \left| \int_0^1 \frac{d\boldsymbol{\gamma}}{d\lambda}(u) du \right| . \tag{2}$$

To enforce a monotonic relationship between high value regions of $\mathcal{E}$ and high importance regions of f, we propose to introduce weights related to the variations of f along the path integral (2). We define these weights (w) at every $u \in [0, 1]$ based on the variations $\dfrac{d(f \circ \boldsymbol{\gamma})}{d\lambda}(u) = \dfrac{\partial f}{\partial \mathbf{x}}(\boldsymbol{\gamma}(u))\dfrac{d\boldsymbol{\gamma}}{d\lambda}(u)$. Several strategies are possible for w. The expressions studied in section 3.2 can be summarized using a continuous function of two variables F, setting $w(u) = F\left(\frac{\partial f}{\partial \mathbf{x}}(\boldsymbol{\gamma}(u)), \frac{d\boldsymbol{\gamma}}{d\lambda}(u)\right)$.

We then define the visual explanation map as

$$\mathcal{E}_w(\mathbf{x}) = \left| \int_0^1 w(u) \frac{d\boldsymbol{\gamma}}{d\lambda}(u) du \right| = \left| \int_0^1 F\left(\frac{\partial f}{\partial \mathbf{x}}(\boldsymbol{\gamma}(u)), \frac{d\boldsymbol{\gamma}}{d\lambda}(u)\right) \frac{d\boldsymbol{\gamma}}{d\lambda}(u) du \right| \tag{3}$$

Weights w, as well as path γ and its derivative, are of the same dimension as image $\mathbf{x}$, summation and multiplication are thus done pixel-wise.

3.2 Choice of the path and regularization

Ideally path $\boldsymbol{\gamma}$ should be traced on the manifold of real clinical images (as in [23]). In practice, this constraint induces heavy computation burdens to determine the derivative $\frac{d\boldsymbol{\gamma}}{d\lambda}$ [3]. To tackle this issue we use a similar expression as [26] and define

[3] As in [3,23], consider an encoder(E)-generator(G) architecture. E (resp. G) maps from (resp. to) the space of real images ($\subset \mathbb{R}^n$) to (resp. from) an encoding space ($\subset \mathbb{R}^k$). The *real* images path γ can for instance be defined as $\gamma : \lambda \to G(z_{\mathbf{x}} + \lambda(z_{\mathbf{x_a}} - z_{\mathbf{x}}))$, where $z_{\mathbf{x}} = E(\mathbf{x})$ and $z_{\mathbf{x_a}} = E(\mathbf{x_a})$. It follows that $\frac{d\gamma}{d\lambda} = \frac{\partial G}{\partial z}(z_{\mathbf{x}} + \lambda(z_{\mathbf{x_a}} - z_{\mathbf{x}}))(z_{\mathbf{x_a}} - z_{\mathbf{x}})$. But $\frac{\partial G}{\partial z}$ is a vector of dimension $n.k$ which easily reaches a magnitude of 10^9 that is to be computed at several values of λ.

$\boldsymbol{\gamma}(\lambda) = \mathbf{x} + \lambda(\mathbf{x_a} - \mathbf{x})$ so that $\frac{d\boldsymbol{\gamma}}{d\lambda} = (\mathbf{x_a} - \mathbf{x})$. Experimentally, even with this simplification, visual explanation maps integrating feature importance (FI)

$$\begin{aligned} \mathcal{E}_{FI}^{v1}(x) &= (\mathbf{x_a} - \mathbf{x})^2 \left| \int_0^1 \frac{\partial f}{\partial \mathbf{x}}(\boldsymbol{\gamma}(u))du \right| \\ \mathcal{E}_{FI}^{v2}(x) &= (\mathbf{x_a} - \mathbf{x})^2 \int_0^1 \left| \frac{\partial f}{\partial \mathbf{x}}(\boldsymbol{\gamma}(u)) \right| du \end{aligned} \tag{4}$$

outperform state-of-the-art methods (section 4). $\mathcal{E}_{FI}^{v1}$ is obtained by setting $F : (x, y) \to x.y$ and is used as baseline. To take into account all derivatives regardless of their signs we set $F : (x, y) \to |x|.y$ and obtain $\mathcal{E}_{FI}^{v2}$. Finally, despite the accumulation of gradients along the linear path between $\mathbf{x}$ and $\mathbf{x}_a$, $\mathcal{E}_{FI}^{v1}$ and $\mathcal{E}_{FI}^{v2}$ tend to be noisy. We thus introduce a regularized version

$$\mathcal{E}_{FI,k_\sigma}^{v2}(x) = \int_0^1 \left((\mathbf{x_a} - \mathbf{x})^2 \left| \frac{\partial f}{\partial \mathbf{x}}(\boldsymbol{\gamma}(u)) \right| \right) * k_\sigma du \tag{5}$$

where k_σ is a centered Gaussian kernel of variance σ. In our experiments $\mathcal{E}_{FI,k_\sigma}^{v2}$ is competitive with $\mathcal{E}_{FI}^{v1}$ and $\mathcal{E}_{FI}^{v2}$ for both the AOPC metric and the feature relevance score on feature importance evaluation [13,17] while improving pathology localization performance.

4 Experiments and Results

4.1 Datasets and Models

Slice classification for brain tumor detection- We use the dataset of Magnetic Resonance Imaging (MRI) for brain tumor segmentation from the Medical Segmentation Decathlon Challenge [22]. Only using the contrasted T1-weighted (T1gd) sequence, we transform the 4-level annotations into binary masks. We then resize the 3D volumes and corresponding binary masks from 155x240x240 to 145x224x224. From these resized masks we affect a class label to each single slice of the volume along the axial axis. Class **1** is given to a slice if at least 10 pixels (0.02 %) are tumorous. Then, the objective is to train a classifier to detect slices with tumors. As additional prepossessing, we remove all slices outside the body along the axial axis and normalize all slice images in [0, 1]. 46900 slices are used for training, 6184 for validation and 9424 for test with 25% of slices with tumors (in each set).

Pneumonia detection- We also use a chest X-Ray dataset from the available RSNA Pneumonia Detection Challenge which consists of X-Ray dicom exams extracted from the NIH CXR14 dataset [27]. We constitute our binary database with 8851 healthy and 6012 pathological exams.As in [4], we only keep the healthy and pathological exams constituting a binary database of 14863 samples (8851 healthy / 6012 pathological). Images are rescaled from 1024 x 1024 to 224 x 224 and normalized to [0, 1]. Bounding box annotations around opacities are provided for pathological cases. The split consists in 11917 exams in train, 1495 in validation and 1451 in test.

Classifier- For the two problems, we train an adapted ResNet50 [11] using the Adam optimizer [12] with an initial learning rate of 1e-4, and minimizing a weighted binary cross-entropy. We recover the ResNet50 backbone trained on ImageNet [6], and add a dropout layer (rate=0.3), then two fully connected layers with respectively 128 and 1 filters. We also introduce random geometric transformations such as flip, rotation, zoom or translation during the training. The classifiers respectively achieve 0.975 and 0.974 AUC scores on brain tumor and pneumonia detection problems.

4.2 Attribution techniques and implementation details

Baseline methods- We compare our method to several baselines and state-of-the-art visual explanation approaches:
(1) Gradient-based (or CAM-based): Gradient [21], Integrated Gradient [26] (IG), GradCAM [19] (GCAM).
(2) Perturbation-based: Mask Perturbation [9] (MPert), Mask Generator [5] (MGen), Similar and Adversarial Generations [4] (SAGen).
(3) Adversary based: the two variations proposed in [3]: CyCE and SyCE. In CyCE, the explanation is computed with the input image $|\mathbf{x} - \mathbf{x_a}|$, while the stable image is used in SyCE version ($|\mathbf{x_s} - \mathbf{x_a}|$).

For MPert [9], we look for a mask of size 56 x 56, and filter it after upsampling ($\sigma = 3$). We use gaussian blur ($\sigma = 5$) to perturbate the input through the mask. Masks are regularized with total variation and finally obtained after 150 iterations. We use a ResNet-50 backbone pre-trained on the task as the encoder part of MGen, then we basically follow the UNet-like [15] architecture and training proposed in [5]. MGen produces mask of size 112x112 that are upsampled to 224x224. For SAGen [4], we use a UNet-like architecture as the common part of the generators and two separated final convolutional blocks for respectively stable and adversarial generations.

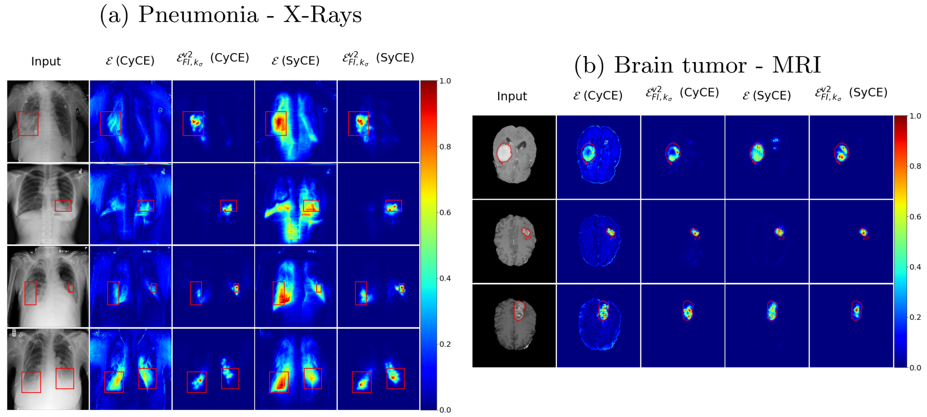
Our methods- For the different variations $\mathcal{E}_{FI}^{v1}$, $\mathcal{E}_{FI}^{v2}$ and $\mathcal{E}_{FI,k_\sigma}^{v2}$, the integral is approximated using a Riemann sum. For instance, in SyCE, $\mathcal{E}_{FI}^{v2}$ is computed through:

$$\mathcal{E}_{FI}^{v2}(x) \approx \frac{(\mathbf{x_s} - \mathbf{x_a})^2}{M} \sum_{m=1}^{M} \left| \frac{\partial f}{\partial \mathbf{x}}(\boldsymbol{\gamma}_m) \right| \tag{6}$$

where $\boldsymbol{\gamma}_m = \mathbf{x_s} + \frac{m-1/2}{M}(\mathbf{x_a} - \mathbf{x_s})$, and M is the number of steps in the Riemann sum. In our experiments, we take $M = 50$. For the regularized version, we apply a Gaussian filtering of kernel 28x28 and σ=2.

4.3 Results

Pathology localization For strong classifiers f, as it is the case here (see sec. 4.1), we expect the feature attributions to match experts annotations as much


(a) Pneumonia - X-Rays

(b) Brain tumor - MRI

Input
$\mathcal{E}$ (CyCE)
$\mathcal{E}^{v2}_{FI,k_\sigma}$ (CyCE)
$\mathcal{E}$ (SyCE)
$\mathcal{E}^{v2}_{FI,k_\sigma}$ (SyCE)
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 1: **Attribution maps visualization**. For (a) and (b), left columns show the input image with red contours indicating pathology localization. Other columns show visualization explanations generated by our regularized method $\mathcal{E}^{v2}_{FI,k_\sigma}$ and two adversarial approaches of [3].

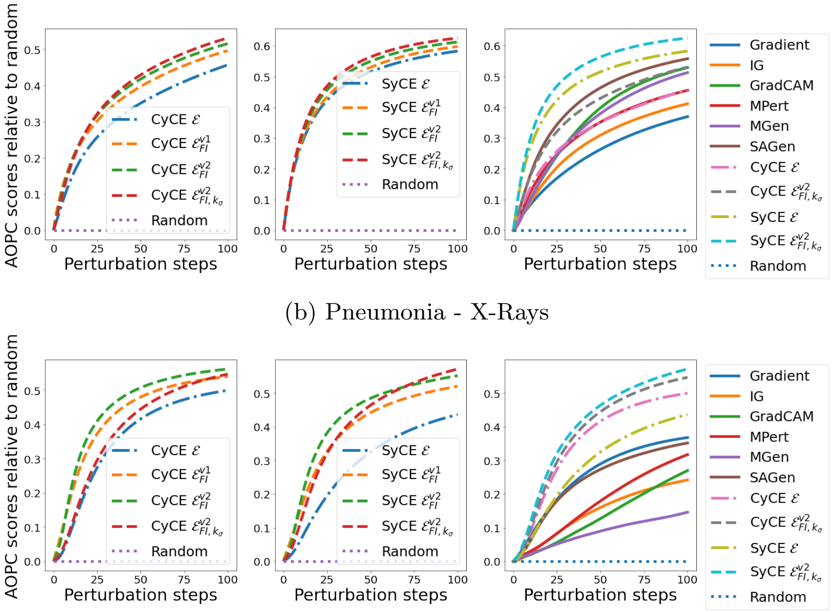
as possible. We measure the intersection over union (IoU)), and the proportion of attribution maps that lays outside the ground truth annotations, also called False Negative Rate (FNR). Both metrics are computed between ground truth annotations and thresholded binary explanation maps. As in [3], the attribution maps are thresholded given percentile values. We only display IoU and FNR scores for percentile values that mostly represent the size distribution of ground truth annotations. Table 1 reports the localization scores. First, adversarial generation methods CyCE and especially SyCE outperform common state-of-the-art approaches. For both SyCE and CyCE, our proposed methods $\mathcal{E}^{v1}_{FI}$, $\mathcal{E}^{v2}_{FI}$ and $\mathcal{E}^{v2}_{FI,k_\sigma}$ significantly improve localization performances compared to the baseline $\mathcal{E}$ (shown in blue or red in table 1), or are at least competitive ($\mathcal{E}^{v1}_{FI}$ for SyCE). In the two adversarial generation approaches, $\mathcal{E}^{v2}_{FI}$ outperforms $\mathcal{E}^{v1}_{FI}$, but the regularized version $\mathcal{E}^{v2}_{FI,k_\sigma}$ is the best localizer (red). Figure 1a and 1b display qualitative results comparing visual explanation from baseline methods CyCE and SyCE with our regularized approach $\mathcal{E}^{v2}_{FI,k_\sigma}$. It visually supports the localization results shown in table 1. Our method focuses only on important region for the classifier which also correlate with human annotations, and remove residual errors remaining in baseline attribution maps (especially for CyCE).

Feature relevance evaluation Although localization performance enables human experts to assess the quality of the visual explanation, it is not enough to translate the importance of features for the classifier. High localization performance does not reflect the capacity of the visual explanation to order regions of the input image w.r.t their importance for the model decision. It only re-

Table 1: **Localization results**. Different attribution methods on Pneumonia detection and Brain tumor problems. IoU (higher is better) and FNR (lower is better) scores are given at representative percentile values for each problem.

	Metric	Perc.	Grad.	IG	GCAM	MPert	MGen	SAGen	CyCE $\mathcal{E}$	CyCE $\mathcal{E}_{FI}^{v1}$	CyCE $\mathcal{E}_{FI}^{v2}$	CyCE $\mathcal{E}_{FI,k_\sigma}^{v2}$	SyCE $\mathcal{E}$	SyCE $\mathcal{E}_{FI}^{v1}$	SyCE $\mathcal{E}_{FI}^{v2}$	SyCE $\mathcal{E}_{FI,k_\sigma}^{v2}$
Pneumonia	IoU ↑	90	0.187	0.170	0.195	0.204	0.208	0.232	0.221	0.269	0.300	**0.321**	0.299	0.289	0.323	**0.335**
		95	0.152	0.136	0.138	0.154	0.169	0.173	0.191	0.232	0.259	**0.278**	0.244	0.242	0.271	**0.285**
		98	0.097	0.086	0.070	0.087	0.103	0.097	0.116	0.149	0.163	**0.175**	0.151	0.150	0.165	**0.177**
	FNR ↓	90	0.639	0.698	0.645	0.623	0.620	0.584	0.596	0.532	0.494	**0.471**	0.492	0.507	0.469	**0.457**
		95	0.584	0.653	0.618	0.576	0.542	0.535	0.494	0.421	0.379	**0.352**	0.399	0.407	0.363	**0.344**
		98	0.508	0.603	0.593	0.537	0.461	0.495	0.430	0.321	0.285	**0.257**	0.319	0.314	0.275	**0.250**
Brain Tumor	IoU ↑	98	0.154	0.238	0.173	0.290	0.318	0.330	0.322	0.337	0.376	**0.428**	0.411	0.404	0.426	**0.432**
		99	0.131	0.196	0.115	0.263	0.274	0.284	0.270	0.288	0.322	**0.363**	0.348	0.338	0.357	**0.364**
	FNR ↓	98	0.744	0.621	0.715	0.580	0.534	0.525	0.542	0.518	0.481	**0.433**	0.462	0.462	0.446	**0.441**
		99	0.687	0.536	0.701	0.451	0.413	0.408	0.440	0.401	0.358	**0.311**	0.344	0.347	0.329	**0.324**

(a) Brain tumor - MRI



(b) Pneumonia - X-Rays

Fig. 2: **AOPC scores relative to random baseline**. The first two columns: baseline adversarial approaches CyCE $\mathcal{E}$ (left) and SyCE $\mathcal{E}$ (middle) compared with our proposed methods. Last column: comparision of baseline $\mathcal{E}$ and regularized version $\mathcal{E}_{FI,k_\sigma}^{v2}$ with other state-of-the-art methods. Results are given for the two classification problems. The higher area, the better.

ports on its capacity to find these regions. To evaluate feature importance for the classifier, we use two metrics based on input degradation techniques [17]: **(i)** the area over the perturbation curve (AOPC) by progressively perturbing the input, starting with the most relevant regions of the explanation map first (introduced in [17]); and **(ii)** the feature relevance score (R) proposed in [13] which combines degradation (most relevant first) and preservation (least relevant first) impacts w.r.t. the classifier. For both a perturbation method must be set. In our experiments, we use an adversarial perturbation (as in [2]). Other types of perturbations (replacement by zero, replacement by noise) generate images outside of the training distribution and break down all visual explanation methods, rendering their evaluation impossible. The adversarial perturbation process for these metrics is independent from adversary generations in adversarial-based visual explanations to produce fair evaluation comparisons. It basically follows the image-to-image translation approach proposed in [20]. The two metrics are computed on a balanced subset of 1000 images of the test set.

Table 2 shows the feature relevance score R for specific (”0” or ”1”) and combined (”All”) predicted classes. Then, figures 2a and 2b show the evolution of the AOPC score on the two classification problems for the different visual explanation approaches relative to a random baseline. (i) Our proposed methods improve adversarial generation baselines CyCE and SyCE for both relevance score on the two predicted classes (blue and red in table 2), and the AOPC metric (red, green and yellow curves compared to the blue one on the first two columns of figures 2a and 2b). (ii) The regularized version $\mathcal{E}^{v2}_{FI,k_\sigma}$ (red curve) is competitive with $\mathcal{E}^{v1}_{FI}$ and $\mathcal{E}^{v2}_{FI}$ (or even outperformed them on Brain tumor problem). (iii) Our methods outperform state-of-the-art approaches (last column in the AOPC figures), especially the ones based on SyCE adversarial generation (see table 2).

Table 2: **Feature Relevance Score** R. Comparing the different attributions methods on Pneumonia detection and Brain tumor problems. The score R is given for specific predicted class 0 and 1 as well as for the two combined (All).

	Pred. Class	Random	Grad.	IG	GCAM	MPert	MGen	SAGen	CyCE				SyCE			
									$\mathcal{E}$	$\mathcal{E}^{v1}_{FI}$	$\mathcal{E}^{v2}_{FI}$	$\mathcal{E}^{v2}_{FI,k_\sigma}$	$\mathcal{E}$	$\mathcal{E}^{v1}_{FI}$	$\mathcal{E}^{v2}_{FI}$	$\mathcal{E}^{v2}_{FI,k_\sigma}$
Pneumonia	All	0.118	0.483	0.385	0.473	0.475	0.339	0.463	0.636	0.672	**0.686**	**0.686**	0.593	0.657	0.679	**0.712**
	0	0.172	0.597	0.562	0.563	0.582	0.304	0.504	0.771	0.830	**0.839**	0.810	0.758	0.823	**0.844**	0.835
	1	0.049	0.037	0.009	0.363	0.255	0.368	0.084	0.357	0.359	0.376	**0.462**	0.298	0.342	0.358	**0.506**
Brain Tumor	All	0.066	0.563	0.599	0.714	0.631	0.689	0.736	0.638	0.675	0.693	**0.703**	0.754	0.766	0.778	**0.788**
	0	0.077	0.509	0.476	0.712	0.460	0.580	0.686	0.486	0.535	0.569	**0.570**	0.702	0.714	0.737	**0.753**
	1	0.055	0.614	0.708	0.715	0.780	0.787	0.783	0.767	0.795	0.801	**0.820**	0.803	0.815	0.817	**0.821**

5 Conclusion

We propose a unification of adversarial visual explanation methods and path-based feature attribution approaches. Using a linear path between the input image and its generated adversary, we introduce a tractable method to assign a weight along this path translating variations of the classifier output. Our method better assesses feature importance attribution compared to both adversarial generation approaches and path-based feature attribution methods. We also improve relevant regions localization performances by reducing the residual reconstruction errors inherent to adversarial generation methods.

References

1. Bien, N., Rajpurkar, P., Ball, R., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B., Yeom, K., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D., Beaulieu, C., Riley, G., Stewart, R., Blankenberg, F., Larson, D., Lungren, M.: Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. PLOS Medicine **15** (2018)
2. Chang, C.H., Creager, E., Goldenberg, A., Duvenaud, D.K.: Explaining image classifiers by counterfactual generation. In: ICLR (2019)
3. Charachon, M., Cournède, P., Hudelot, C., Ardon, R.: Leveraging conditional generative models in a general explanation framework of classifier decisions. In: ArXiv (2021)
4. Charachon, M., Hudelot, C., Cournède, P.H., Ruppli, C., Ardon, R.: Combining similarity and adversarial learning to generate visual explanation: Application to medical image classification. In: ICPR (2020)
5. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. In: NIPS (2017)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
7. Elliott, A., Law, S., Russell, C.: Adversarial perturbations on the perceptual ball. ArXiv **abs/1912.09405** (2019)
8. Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. In: Nature. vol. 542 (2017)
9. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: ICCV (2017)
10. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: ICML (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
13. Lim, D., Lee, H., Kim, S.: Building reliable explanations of unreliable neural networks: Locally smoothing perspective of model interpretation. In: CVPR (2021)
14. Pratt, H., Coenen, F., Broadbent, D.M., Harding, S.P., Zheng, Y.: Convolutional neural networks for diabetic retinopathy. Procedia Computer Science **90**, 200–205 (2016). https://doi.org/https://doi.org/10.1016/j.procs.2016.07.014, `https://www.sciencedirect.com/science/article/pii/S1877050916311929`, 20th Conference on Medical Image Understanding and Analysis (MIUA 2016)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
16. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
17. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.: Evaluating the visualization of what a deep neural network has learned. In: IEEE Transactions on Neural Networks and Learning Systems (2017)
18. Seah, J.C.Y., Tang, h.J.S.N., Kitchen, A., Gaillard, F., Dixon, A.F.: Chest Radiographs in Congestive Heart Failure: Visualizing Neural Network Learning. In: Radiology. vol. 290

19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: ICCV (2017)
20. Siddiquee, M.R., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M., Bengio, Y., Liang, J.: Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: ICCV. pp. 191–200 (2019)
21. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: ICLR (2014)
22. Simpson, A., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Ginneken, B., Kopp-Schneider, A., Landman, B., Litjens, G., Menze, B., Ronneberger, O., Summers, R., Bilic, P., Christ, P., Do, R., Gollub, M., Golia-Pernicka, J., Heckers, S., Jarnagin, W., McHugo, M., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. ArXiv **abs/1902.09063** (2019)
23. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration. In: ICLR (2020)
24. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. ArXiv **abs/1706.03825** (2017)
25. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. In: ICLR. vol. abs/1412.6806 (2015)
26. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML (2017)
27. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR (2017)
28. Woods, W., Chen, J., Teuscher, C.: Adversarial explanations for understanding image classification decisions and improved neural network robustness. In: Nature Machine Intelligence. vol. 1 (2019)
29. Xu, S.Z., Venugopalan, S., Sundararajan, M.: Attribution in scale and space. In: CVPR (2020)
30. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)
31. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)